НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

«На правах рукопису»
УДК 004.855.2

«До захисту допущено»
Завідувач кафедри
_____ О.Л.Тимощук
«___»_____20__ р.

Магістерська дисертація
на здобуття ступеня магістра
зі спеціальності 124 Системний аналіз
на тему: «Система аналізу неструктурованих текстових даних»

Виконав:
студент ІІ курсу, групи КА-61м
Шибирин Ігор Олегович _____

Керівник:
старший науковий співробітник
ІПС НАН України, к.ф.-м.н, с. н. с.
Ігнатенко О. П. _____

Рецензент:
старший науковий співробітник відділу
Нейротехнологій ІПММС НАН України, к.т.н., с.н.с.
Чернодуб А.М. _____

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студент _____

Київ
2018

РЕФЕРАТ

Магістерська дисертація: 123 с., 15 рис., 25 табл., 2 додатки, 25 джерел.

Об’єкт дослідження – складні системи аналізу неструктурованих текстових даних.

Предмет дослідження – методи та системи аналізу неструктурованих текстових даних.

Мета роботи – розробка та дослідження автоматичної системи визначення настроїв щодо аспектів продукту, що дає можливість усувати інформаційну невизначеність при прийнятті управлінських рішень.

Наукова новизна роботи визначається наступним теоретичними і практичними результатами, отриманими автором:

- уперше запропоновано використовувати особливості домену в детерміністичному підході для визначення думки щодо аспектів продукту;
- уперше виконано програмну реалізацію модифікованого методу детерміністичного підходу для визначення думки щодо аспектів продукту.

Результати даної роботи рекомендується використовувати для розробки складної інтелектуальної системи прийняття рішень, що базуються на відгуках клієнтів.

СИСТЕМИ АНАЛІЗУ НЕСТРУКТУРОВАНИХ ТЕКСТОВИХ ДАНИХ, СИСТЕМА АНАЛІЗУ НАСТРОЮ ТЕКСТУ, КЛАСИФІКАЦІЯ, АСПЕКТНО-ОРІЄНТОВАНИЙ АНАЛІЗ ТЕКСТУ, ВИЗНАЧЕННЯ АСПЕКТІВ.

ABSTRACT

Master's dissertation: 123 pages, 15 figures, 25 tables, 25 sources.

The object of the study - complex systems for analyzing unstructured text data.

Subject of research – methods and systems for analyzing unstructured text data.

Purpose of work – research and development of an automatic system for sentiment analysis in relation to product aspects, which makes it possible to eliminate information uncertainty when making managerial decisions.

The purpose of the work is the development and research of an automatic system for sentiment analysis in relation to product aspects, which makes it possible to eliminate information uncertainty when making managerial decisions.

The scientific novelty of the work is determined by the following theoretical and practical results obtained by the author:

- for the first time, it is proposed to use domain features in the deterministic approach to determine the views of product aspects;
- for the first time a programmatic implementation of the modified method of deterministic approach has been implemented to determine the views on product aspects.

The results of this work are recommended for developing a sophisticated intellectual decision-making system based on customer feedback.

SYSTEMS OF ANALYSIS OF UNSTRUCTURED TEXT DATA, TEXT SETTING ANALYSIS SYSTEM, CLASSIFICATION, ASPECT-ORIENTED ANALYSIS OF TEXT, ASPECT MINING.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

BD – Big Data

SA – Sentiment Analysis

ML – Machine Learning

BN – Bayesian Network

NN – Neural Network

SVM – Supporting Vector Machine

BOW – Bag of Words

PMI – Point Mutual Information

FS – Features Selection

BNS – Bi-Normal Separation

LSI – Latent Semantic Indexing

POS – Point of Speech

ME – Maximum Entropy

MST Maximum Spanning Tree

DT Decision Trees

ВСТУП

Кількість даних, що потребують аналізу зростають кожного дня. Це породжує багато проблем, таких як: обробка та аналіз даних, швидкість аналізу, моніторинг даних. Для вирішення цих проблем були розроблені технології великих даних Big Data - потужний інстумент в руках системного аналітика. Вони допомагають проаналізувати величезні обсяги даних, щоб знайти залежності між елементами системи, знайти сенс в до того протирічних даних, зменшити невизначеність. Аналіз великих даних дозволяє отримати більш повний погляд на прикладні проблеми.

В останні роки було розроблено велику кількість алгоритмів та аналітичних рішень що використовують дані користувача (наприклад вік, стать, статистику використання сервісом тощо). Більшість даних, що оброблюються, мають просту природу - число, дата, належність певній групі ознак. Для таких даних розроблені ефективні засоби аналізу.

Проте є значно складніший тип даних - наприклад текст, відео, аудіо. Ці дані називають неструктурованими і вони представляють особливий інтерес для аналізу, оскільки характеризуються неповнотою, протирічністю та невизначеністю. Саме такі характеристики має задача системного аналізу.

Для розв’язання цієї задачі необхідно розробити систему аналізу, яка здатна в режимі реального часу оброблювати велику кількість неструктурованих текстових даних.

Ця задача є особливо цікавою, оскільки саме в такому вигляді люди звикли ділитися своїми думками та враженнями у соціальних мережах, відгуках, блогах тощо.

Прикладом задач які вирішуватиме система є аналіз відгуків про продукти - він допомагає знайти основні характеристики в яких зацікавлені користувачі. Згодом цю інформацію можна буде використати під час наступних стадій аналізу або прийняття рішень - наприклад морфологічного аналізу. Відгуки про сервіси дозволяють зрозуміти потреби користувача, які послуги можна покращити, які нові послуги або продукти можна створити та на що звернути увагу під час роботи організації.

Додаткова інформація отримана завдяки аналізу великих даних збільшує прибутки, створює кращі товари та послуги.

Саме з цих причин ціль роботи заключається в розробці системи аналізу неструктурованих текстових даних.

В ході роботи були вирішені наступні задачі:

- аналіз алгоритмів та підходів до визначення настрою тексту;
- аналіз алгоритмів та підходів до визначення аспектів продукту про які було згадано в тексті;
- розробка методу визначення аспектів та відношення автора до них;
- розробка візуального представлення результатів роботи методу.

Результат для практичного застосування - програмна реалізація запропонованого методу, що використовує мови програмування Python та JavaScript.

РОЗДІЛ 1. ОГЛЯД ПОТОЧНОГО СТАНУ ПРОБЛЕМИ

1.1 Аналіз актуальності задачі аналізу великих неструктурованих текстових даних

Концепція Big Data існує протягом багатьох років; більшість організацій сьогодні розуміють, що якщо вони зберігають всі дані, що отримують впродовж життя свого бізнесу, вони можуть застосувати аналітику та отримати з неї значну користь. Але навіть у 1950-х роках, за десятиліттями, ще до того коли хтось перший використав термін "великі дані", бізнес використовував базову аналітику (по суті числа в електронній таблиці, які аналітик перевіряв вручну), щоб розкрити закономірності, взаємозв’язки та тенденції.

Однак переваги, які аналітика великих даних приносить в аналіз - це швидкість та ефективність. Тоді як кілька років тому бізнес міг би зібрати інформацію, залучити аналітика та розкрити тенденції, які можна використати для майбутніх рішень, сьогодні бізнес може отримати поради щодо негайних дій. Здатність працювати швидше - і залишатися гнучкими - надає організаціям конкурентну перевагу, яку вони не мали раніше.

Швидкість обробки особливо важлива у зв’язку зі зростаючою тенденцією по накопиченню даних. Згідно з оцінками, до 2009 року майже всі сектори економіки США мали в середньому 200 терабайт збережених даних (обсяг удвічі більший за складську базу даних найбільшої мережі роздрібних магазинів в США - Wal-Mart - станом на 1999 рік) для компаній з більш ніж 1000 працівниками.

За іншими оцінками очікується значне зростання потреби в аналітиках та програмних комплексах що здатні працювати з великими даними. Так відсоток даних, які будуть корисними для аналізу зросте з 22% до більш ніж 35%. Очікується, що всесвітній ринок великих даних (технологій та послуг з їх обробки) зростатиме зі швидкістю близько 23% щорічно в період між 2014 і 2019 роками, а світові доходи від великих даних та бізнес-аналізу збільшиться більш ніж на 50% з майже 122 мільярдів доларів США у 2015 році до більш ніж 187 мільярдів доларів США в 2019 році. Найбільші сектори в яких використовується BD

включають виробництво, банківська справа та страхування, телекомунікації, охорону здоров'я, транспорт та роздрібну торгівлю.

Аналітика великих даних допомагає організаціям використовувати впродовж своєї роботи дані для виявлення нових можливостей. Це, в свою чергу, призводить до прийняття більш інформованих бізнес рішень, проведення ефективніших операцій, отримання вищого прибутку та щасливих клієнтів. У своєму дослідженні "Великі Дані у великих компаніях"[1], директор з досліджень IIA Том Давенпорт розглянув більше 50 підприємств, щоб зрозуміти, як вони використовували великі дані. Він виявив, що BD приносять користь такими способами:

- зниження вартості - Hadoop та хмарна аналітика, приносять значні переваги при зберіганні великої кількості даних, а також можуть виявити більш ефективні способи ведення бізнесу;
- швидше, краще прийняття рішень - завдяки швидкості роботи Hadoop та аналізі пам'яті в поєднанні з можливістю аналізу нових джерел даних підприємства можуть негайно аналізувати інформацію та приймати рішення на підставі того, що вони дізналися;
- нові продукти та послуги - завдяки здатності оцінювати потреби клієнтів та їх задоволення від продукту з’являється можливість надати клієнтам те, що вони хочуть.

Окрім вигоди для бізнесу, аналіз великих даних може бути застосований і в інших областях. За його допомогою можна оцінювати якість освіти, глобальну зміну клімату, глобальні тенденції щодо зайнятості, екологічну ситуацію тощо.

1.2 Цілі аналізу великих даних

Як і багато нових інформаційних технологій, великі дані можуть призвести до значного скорочення часу, необхідного для обчислень, або створення нових продуктів та послуг.

Як і традиційна аналітика, BD здатна підтримувати внутрішні бізнес-рішення. Технології та концепції на яких будується аналіз великих даних організації дозволяють досягати поставлених цілей.

1.2.1 Зниження вартості за допомогою BD

Організації, орієнтовані на скорочення витрат, прийняли рішення прийняти інструменти великих даних для роботи з даними в межах своїх підрозділів інформаційних технологій керуючись техніко-економічним критерієм.

Зниження вартості може бути додатковою метою після досягнення інших цілей. Скажімо, перша мета організації - інновації в продуктах та послугах, що були отримані за допомогою BD.

Після досягнення цієї мети вона може захотіти зменшити витрати.

1.2.2 Зменшення часу

Другою метою BD є скорочення часу. Мережа універмагів Macy`s змогла скоротити час на ціноутворення для 73 мільйонів одиниць товару з 27 годин до 1 години. Ця можливість дозволяє їй проводити переоцінку

товарів набагато частіше й адаптуватися до зміни умов на роздрібному ринку. Розроблена система аналізу бере дані з Hadoop кластера та поміщає його в програмне забезпечення що дозволяє розпаралелити обчислення. Масу`s стверджує що завдяки цьому витрати на апаратне забезпечення були скорочені на 70%.

Ще однією ключовою метою є можливість взаємодії з клієнтом у режимі реального часу, використовуючи аналітику та дані, отримані від досвіду клієнтів. Аналіз зворотнього зв'язку допомагає виправити недоліки власного продукту та уникнути помилок, зроблених конкурентами.

1.2.3 Створення нових пропозицій

Одна з найамбіційніших задач, яку можна розв'язати за допомогою великих даних - використати їх розробки нових продуктів та послуг. Багато компаній, які використовують цей підхід - онлайн-фірми, що отримують прибуток від продуктів та послуг пов'язаних з даними. Яскравим прикладом може бути LinkedIn, який використовував BD для розробки широкого спектру пропозиції продуктів і нових сервісів, у тому числі люди, яких ви можете знати, групи, які можуть вам сподобатися, хто переглядав мій профіль та інші. Ці сервіси допомогли завоювати мільйони нових клієнтів.

1.2.4 Підтримка внутрішніх бізнес-рішень

Основна мета традиційної аналітики так званих "малих даних" - підтримка внутрішніх бізнес-рішень Які пропозиції цікавлять клієнта? Хто перестане користуватися сервісом найближчим часом? Який об’єм товару необхідно утримувати на складі? Яку ціну можна встановити на свій товар?

Ці типи рішень використовують великі дані, якщо є нові, неструктуровані джерела даних, що можуть допомогти знайти рішення. Наприклад, будь-які дані, які можуть допомогти дізнатися чи задоволені клієнти допомагають у прийнятті рішень. Більшість з них - неструктуровані текстові дані.

1.2.5 Збільшення прибутків

За результатами дослідження виконаного MGI та McKinsey's Business Technology Office роздрібні продавці, які використовують великі дані, можуть збільшити свою операційну маржу більш ніж на 60 відсотків. Використання великих даних має величезний потенціал у державному секторі. Наприклад якби система американської охорони здоров'я творчо і ефективно використовувала великі дані для підвищення ефективності та якості, цей сектор зміг би заробляти більше 300 мільярдів доларів щороку. Дві третини з них будуть у вигляді зменшення витрат на охорону здоров'я в США приблизно на 8 відсотків.

1.3 Складові технології аналізу великих даних

Кожен компонент стеку технологій аналізу великих даних оптимізований навколо великого, неструктурованого та напівструктурованого характеру BD.

1.3.1 Сховище

Зберігання великих обсягів різноманітних даних на диску стає більш економічно ефективним, оскільки вартість зберігання зменшується і потреба в даних, які до того вважалися непотрібними, може виникнути в будь-який момент (ці дані можуть принести прибутки в майбутньому).

1.3.2 Інфраструктура платформи

Велика "платформа" даних - це, як правило, сукупність функцій, які відповідають за високу продуктивність обробки. Платформа включає в себе можливості інтегрувати, керувати та застосовувати складні обчислювальні алгоритми для обробки даних. Як правило, великі платформи даних включають Hadoop (або подібний проект з відкритим

кодом). Hadoop був спроектований і побудований для оптимізації складних маніпуляцій коли великі коли дані значно перевищують ціну або продуктивність традиційних баз даних. Hadoop це уніфіковане середовище зберігання та обробки, яке гарно масштабується до великих і складних об'ємів даних.

1.3.3 Дані

Розміри великих даних настільки ж широкі та складні, як і їх застосування. Великі дані можуть означати послідовності геномів людини, датчики нафтових свердловин, поведінку ракових клітин, розташування продуктів на піддонах, взаємодії у соціальних мережах або життєві ознаки пацієнта тощо. Рівень даних в стеку означає, що дані є окремим активом, що гарантує дискретне управління та контроль. З цією метою опитування фахівців з управління даними у 2013 році виявило, що з 339 компаній 71 відсоток відповів, що "ще не планують" свої стратегії передачі BD. Респонденти виявили стурбованість щодо якості даних, своєчасності та безпеки як значущих бар'єрів для прийняття великих даних та їх широкого застосування. Приклад розмірів даних для Британії наведено на рисунку 1.1.

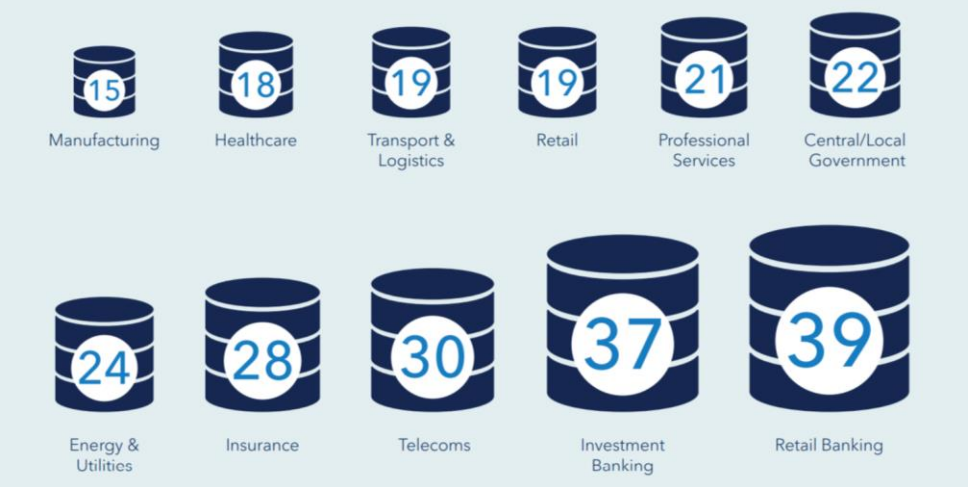

Рисунок 1.1 - Приклад розмірів даних для Британії за останні 12 місяців 2013 року

1.3.4 Код програми, функції та служби

Подібно до того, як BD залежать від бізнес-задачі, код, що використовується для обробки даних може змінюватися. Hadoop використовує процесор під назвою MapReduce, щоб не тільки розподіляти дані по дисках, але й застосовувати складні обчислювальні інструкції до цих них. Відповідно до високої продуктивної можливості платформи, інструкції MapReduce обробляються паралельно через різні вузли на великій платформі даних, а потім швидко зібрані, щоб забезпечити нову структуру даних або відповідь.

Прикладом застосування великих даних у Hadoop може бути "обчислити всіх клієнтів, які нас люблять за допомогою інформації з засобів масової інформації".

1.3.5 Бізнес-логіка (Business View)

Залежно від застосування, можна зробити додаткову обробку через MapReduce або користувацький код на Java для побудови проміжної структури даних, наприклад статистичної моделі, плоского файлу, реляційної таблиці. Отримана структура далі може бути використана для додаткового аналізу або для запиту в традиційних інструментах на базі SQL.

1.3.6 Презентація

Засоби візуалізації даних дозволяють звичайному фахівцю переглядати інформацію в інтуїтивно зрозумілій, графічній формі.

Наприклад, якщо постачальник бездротової мережі бажає отримати кращу інформацію про свою мережу, зокрема, знайти закономірність у скинутих дзвінках, він може зібрати складну електронну таблицю з різними стовпцями та цифрами. З іншого боку він може розгорнути простий споживчий звіт про тенденції, як показано на рисунку 1.2.

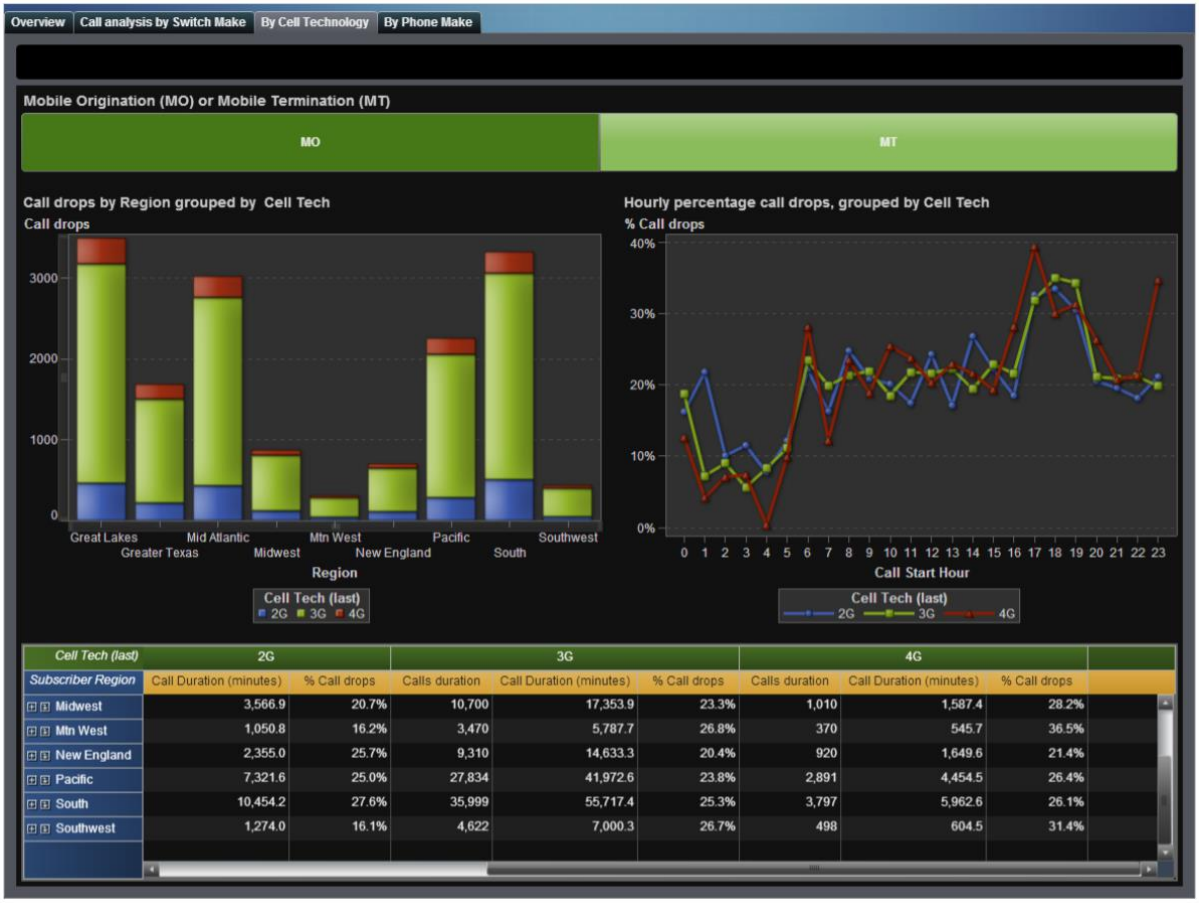

Рисунок 1.2 - Приклад користувацького інтерфейсу

Ця візуалізація даних відображає три різні представлення даних. Перший показує кількість дзвінків по регіонах згруповані за генерацією мережі. Другий показує розподіл скинутих дзвінків у часі. Третій показує більший відсоток випадаючих дзвінків у мережі 4G на початку виклику о 17:00. Така інформація може підказати оператору мережі продовжувати роботу та виявити основні причини проблем в мережі через які особливо цінні клієнти можуть не отримати якісний сервіс.

Таку візуалізацію звичайний оператор мережі здатний завантажити на свій настільний ПК або скинути на мобільний пристрій сервісному техніку в “полях”.

Візуалізація даних зазвичай дуже приваблива для керівників, але чим більша розмірність даних - тим складніше її створювати і сприймати.

Людям складно розібратися у візуалізаціях що мають більше ніж два виміри. Деякі інструменти візуалізації даних вміють обирати найвідповідніший візуальний показ в залежності від типу даних та кількості змінних.

1.4 Результати застосування

Згідно опитування проведеного Ренді Біном[2] проведеного серед президентів, начальників відділів інформації, керівників відділів аналізу, начальників відділів маркетингу які представляють 50 гігантів промисловості, включаючи American Express, Capital One, Disney, Ford Motors, General Electric, JP Morgan, MetLife, Nielsen, Turner Broadcasting, United Parcel Service та USAA, 48.4% з опитуваних досягли результатів, що можна вважати значущими. 80.7% респондентів характеризують свої інвестиції у великі дані як успішні. На рисунку 1.3 зображено фактори, які покращилися внаслідок використання BD.

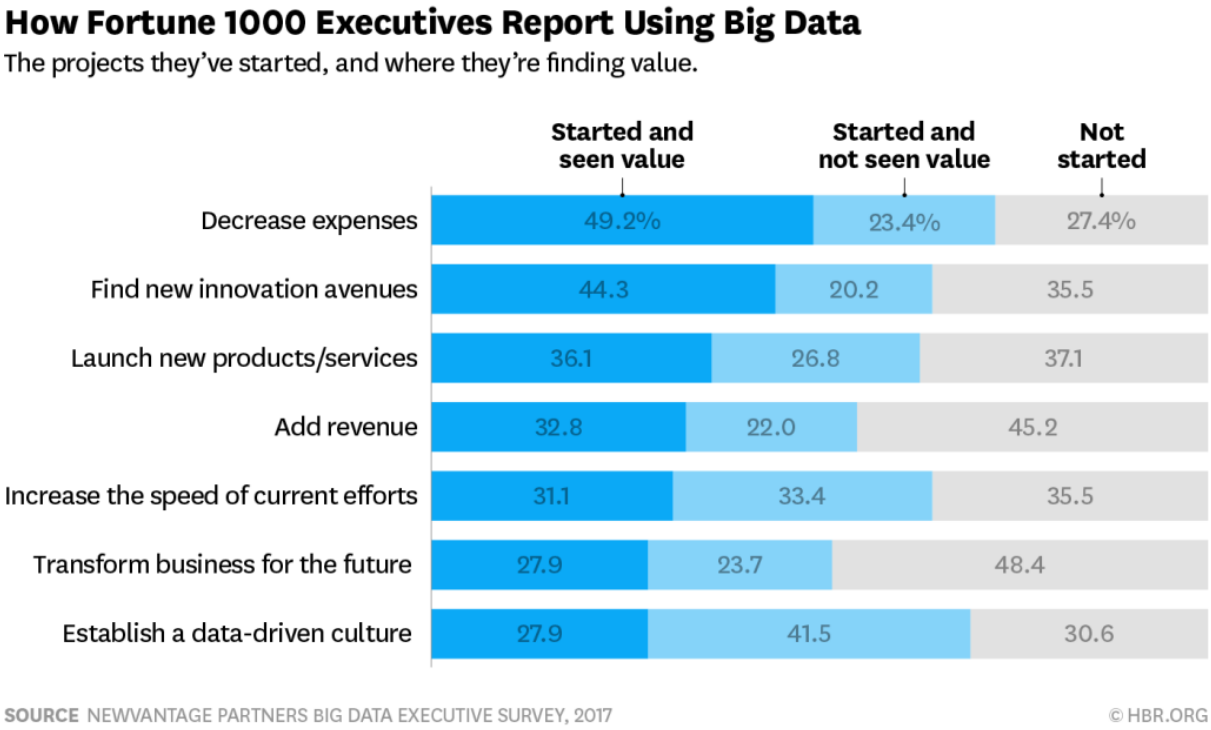

Рисунок 1.3 - Відповіді респондентів щодо користі від великих даних

Роб Петерсен зібрав приклади вдалого використання великих даних[3]. Серед них найцікавішими досягненнями є:

- American Express розробила складну модель з 115 змінними для передбачення вірності своїх клієнтів. Компанія заявляє що здатна визначити близько 24% користувачів що збираються піти впродовж наступних чотирьох місяців.
- Bank of America створює кеш-бек пропозиції для своїх клієнтів що користуються кредитними та дебетовами картами базуючись на їх попередніх покупках.
- Caesars Entartainment поєднує результати азартних ігор споживачів з програмою винагороди, щоб запропонувати привабливі пільги тим, хто програє.
- Duetto персоналізує ціну на номер для людей, які шукають в Інтернеті готелі. Ціни в готелі можуть бути персоналізовані,

враховуючи дані, такі як скільки клієнт витрачає в барі чи казино. Готель може знизити ціну, знаючи, що він заробить гроші на інших послугах.

- Evolv допомагає великим світовим компаніям приймати кращі рішення щодо прийняття управлінських рішень використовуючи прогнозну аналітику. Evolv розглядає понад 500 змінних, таких як дані про ціни на газ, рівень безробіття та використання соціальних мереж. Це допомогає клієнтам, такими як Xerox прогнозувати, коли працівник, найімовірніше, залишить роботу. Компанії, такі як Xerox, AT&T та Kelly Services, використовують Evolv, і в середньому бачать заощадження ресурсів на суму 10 мільйонів доларів. Продажі послуг Evolv зросли на 150% в період з третього кварталу 2012 року до третього кварталу 2013 року.
- General Electric використовує дані записані при використанні техніки (починаючи від локомотивів і закінчуючи медичним обладнанням) для оптимізації їх роботи. Аналітики General Electric передбачають що дані зможуть підвищити продуктивність в США на 1.5%, що за 20 річний період допоможе зберегти достатньо коштів щоб збільшити середні доходи громадян на 30%.
- Google співпрацюючи з Центром по контролю захворювань США відслідковує пошукові запити користувачів щодо симптомів грипу, що допомагає передбачити осередки епідемії.
- Homer завдяки провідним фахівцям у галузі грамотності допомагає дітям навчитися читати. Він має повну фонетичну бібліотеку, ілюстровані оповідання та цікаві предмети мистецтва. Homer поєднує найкращі методи раннього навчання з цікавим додатком, який пов'язує навчання з читанням, щоб навчитися розуміти світ.

- IRS дикористовує великі дані, щоб запобігти крадіжці ідентичності, шахрайству та неналежним платежам, знаходжуючи тих, хто, наприклад, не сплачує податки або штрафи. Система також допомагає забезпечити дотримання податкових правил та законів. Поки що IRS запобіг втраті мільярдів доларів через шахрайство, і за останні три роки повернув понад 2 мільярдів доларів.
- Netflix забезпечив собі велику користувацьку базу за рахунок свого високоякісного контенту. Тепер Netflix використовує свої дані та аналітику про міжнародні звички перегляду, щоб створювати та купувати контент, який, на його думку, охопить великі, сформовані аудиторії телеглядачів.
- Procter & Gamble: вивчення успішності своєї бізнес програми та швидшого реагування на зміни ринкових умов, P&G потрібно було чітко та легко опрацювати свою швидко зростаючу та величезну кількість даних. інтегрувати величезні обсяги структурованих та неструктурованих даних в рамках досліджень і розробок, ланцюжка поставок, операцій із клієнтами та взаємодії з клієнтами, як з традиційних джерел даних, так і з нових джерел онлайнових даних. Тепер P&G може завантажувати та інтегрувати дані швидше і виконувати надійний аналіз обсягів даних, які раніше були неможливими.
- Sprint використовує аналітику великих даних для покращення якості досвіду користувачів, одночасно знижуючи кількість помилок мережі та відтоку клієнтів. Вони обробляють 10 мільярдів транзакцій щодня для 53 мільйонів користувачів, а їх аналітика великих даних допомогла підвищити пропускну здатність на 90%.
- Uber знизила кількість автомобілів на дорогах Лондона на третину завдяки своєму сервісу UberPool, що обслуговує користувачів, які зацікавлені в зниженні їх витрат на паливо. Бізнес

Uber побудований на BD: він використовує дані водіїв та пасажирів, які згодом обробляються алгоритмами, що знаходять підходящий тариф на проїзд.

- Щоденно UPS доставляє 16,9 мільйонів пакунків і документів, щорічно близько 4 мільярдів доставок робляться за допомогою майже 100 000 автомобілів. Телематична техніка та розширені алгоритми допомагають з маршрутами, простою в двигуні та прогнозуванням поломок. З моменту запуску програми компанія заощадила більше 39 мільйонів галонів пального і уникала руху 364 мільйонів кілометрів.

Ще одним доказом того, наскільки сильно Big Data може вплинути на економіку країни слугує звіт компанії Cerb (рисунок 1.4) про значення великих даних в економіці Британії [4].

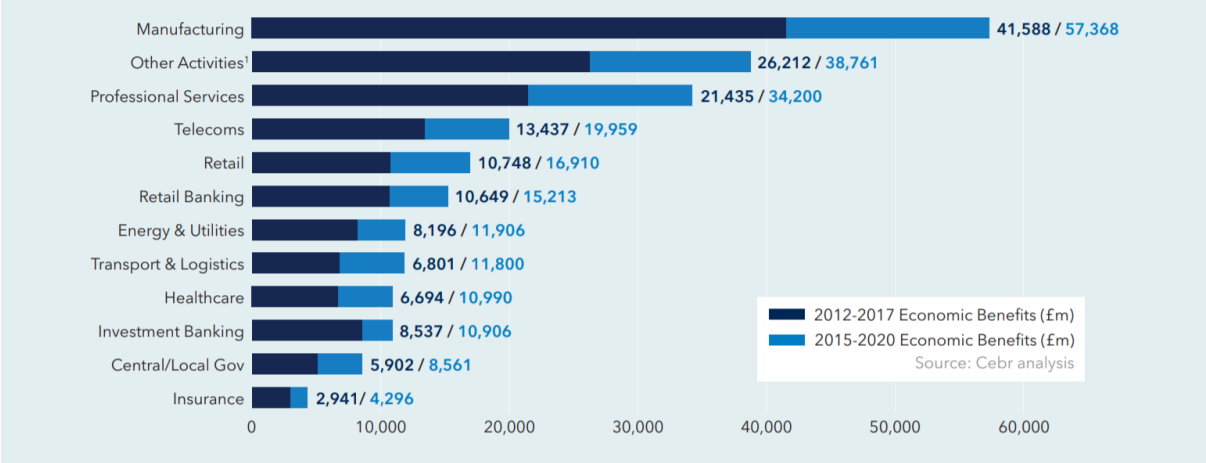

Рисунок 1.4 - Сумарний вплив BD за 2012-2017 та прогнозований вплив з 2015-2020 на економіку Британії поіндустрійно

Передбачається комбінований економічний ефект у розмірі від 322 мільярдів фунтів стерлінгів у період між 2015 і 2020 роками завдяки аналізу великих даних. Таблиця 1.1 показує, що в період між 2012 та 2017 роками очікується, що вигода від великих даних та аналітичних рішень,

які розкривають їх, складають 162 мільярди фунтів стерлінгів або в середньому 27 мільярдів доларів на рік. Це приблизно 1,4% річного ВВП. З 2015 року по 2020 рік оцінка загальної користі від аналізу великих даних для економіки Великої Британії, на суму 241 мільярдів фунтів стерлінгів або в середньому 40 мільярдів фунтів стерлінгів на рік. Це еквівалентно в середньому 2,0% ВВП. Зростання значення великих аналітичних даних в часі є функцією від зростання використання великих даних у різних галузях. Оскільки все більше підприємств користуються аналітикою великих даних, накопичені прибутки, отримані за допомогою електронної комерції, створюються інновації та створення бізнесу. Очікується що до 2020 року вартість аналітики великих даних досягне 46 мільярдів фунтів стерлінгів (ціни 2015 року) або 2,2% ВВП. Таблиця 1.1 також показує механізми, за допомогою яких накопичуються загальні переваги від такої аналітики.

Таблиця 1.1 Загальний вплив від BD (джерело Cerb analysis)

	2012-2017 Прибуток у мільйонах фунтів стерлінгів	2015-2020 Прибуток у мільйонах фунтів стерлінгів
Прибуток від ефективності	145,521	220,373
Прибуток від інновацій	8,341	12,416
Прибуток від створення	8,470	8,082
Загальний прибуток	162,331	240,870

Галузь яка, як очікується, отримає найбільшу економічну вигоду від великих даних - виробництво. Очікується, що загальна вартість великих даних для неї буде збільшена до 57 мільярдів фунтів стерлінгів у 2020 році. Це може бути пов'язано з різноманітністю фірм у галузі та різноманітністю областей, в яких можна досягти високих результатів за

рахунок використання великих даних та аналізу великих даних, таких як поліпшення управління ланцюжком поставок та аналізу клієнтської інформації. На рисунках 1.5-1.7 зображені графіки вигоди від застосування великих даних.

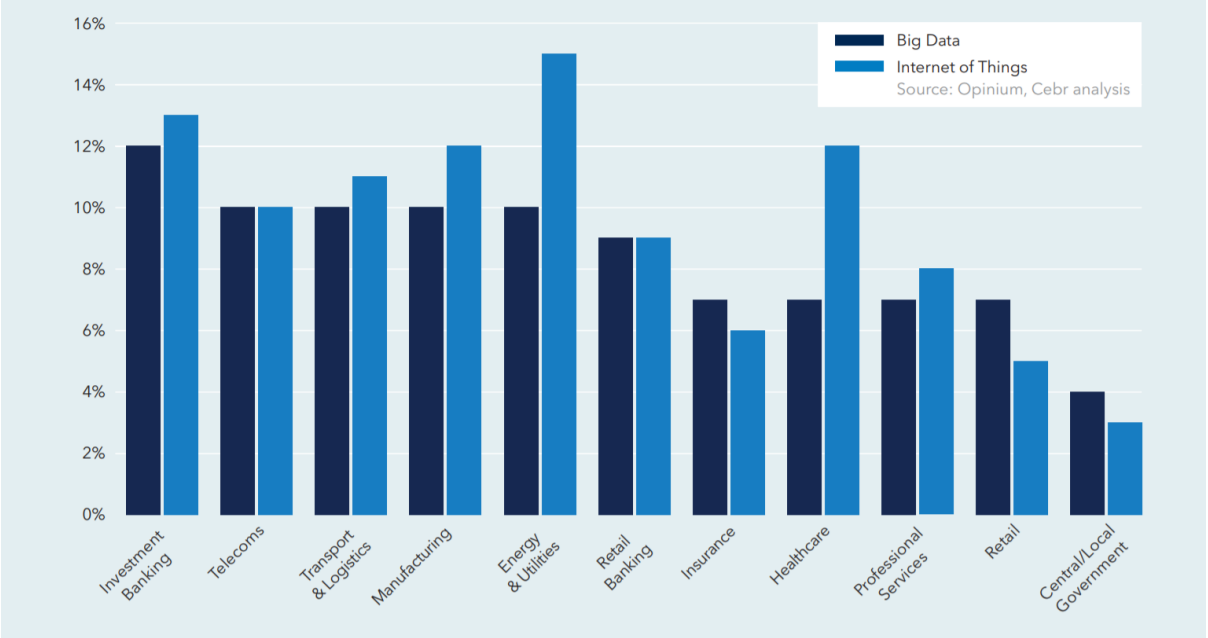

Рисунок 1.5 - Збільшення прибутків поіндустрійно (джерело Opinium, Cerb analysis)

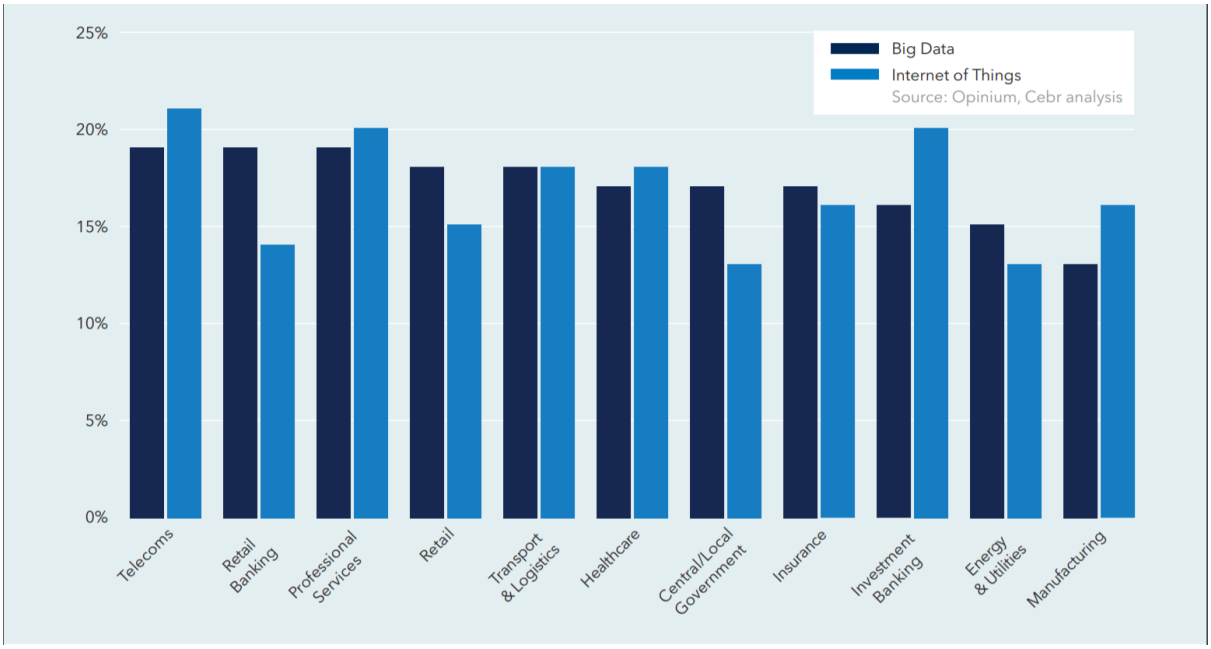

Рисунок 1.6 - Збільшення кількості збережених коштів поіндустрійно (джерело Opinium, Cerb analysis)

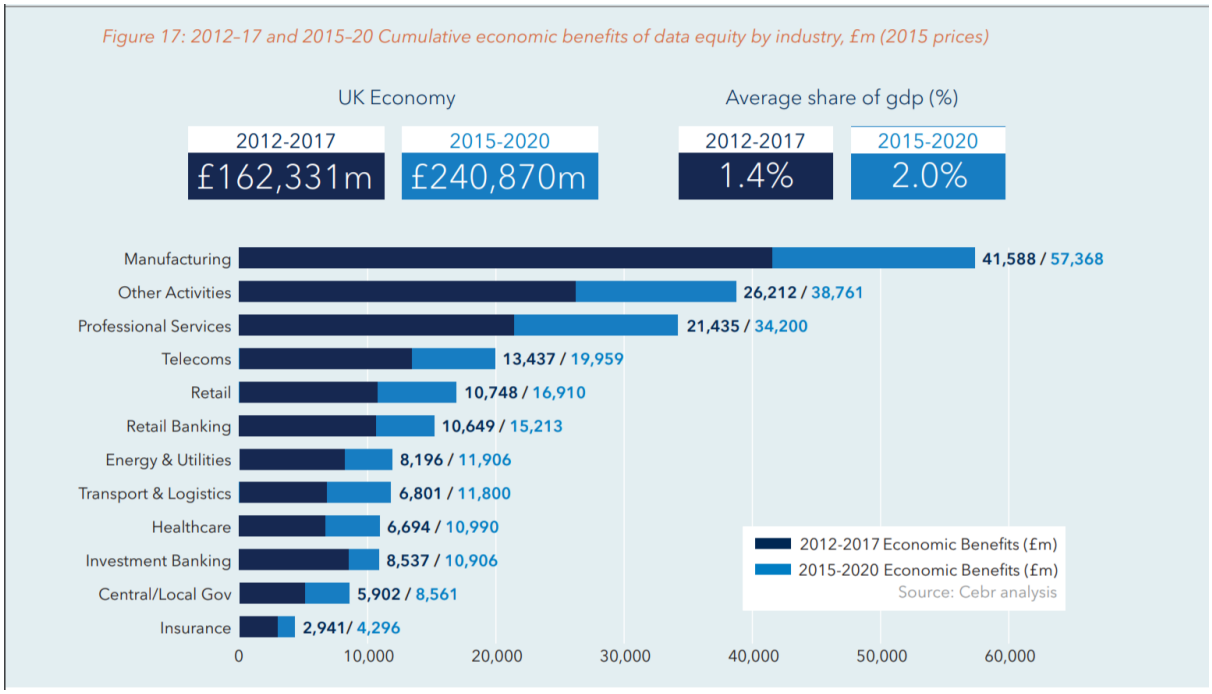

Рисунок 1.7 - Загальний вплив на економіку з 2012-2017 та прогнозований на період з 2015-2020

Також покращиться управління ризиками оскільки з’явиться нова інформація отримана з аналізу, завдяки якій можна буде приймати більш інформовані рішення.

1.5 Великі дані в освітній системі

Користь від аналізу великих даних не обмежується бізнесом, банківською справою та сферою охорони здоров’я. Освіта - ще один перспективний напрямок для використання [5].

1.5.1 Покращення результатів учнів

Загальна мета BD в освітній системі полягає у поліпшенні результатів навчання. Освічені студенти покращують суспільство, корисні для організацій, а також навчальних закладів. Наразі єдиний спосіб оцінки здібностей студента - виконані завдання та іспити. Протягом свого студентського життя кожен створює унікальний слід з даних. Цей слід може бути проаналізованим в реальному часі, щоб забезпечити оптимальне навчальне середовище для студента, а також для кращого розуміння індивідуальної поведінки студентів.

Можна відстежувати кожну дію студентів. Скільки часу вони витрачають, щоб відповісти на питання, які джерела вони використовують, які питання вони пропускали, які поради найкраще працюють для яких студентів тощо. Відповіді на питання можна перевірити миттєво і автоматично (за винятком відкритих питань) та отримувати миттєвий відгук від студентів.

Крім того, BD може допомогти створити групу студентів, які виконують завдання краще. Студенти часто працюють у групах, де студенти не доповнюють одне одного. За допомогою алгоритмів можна буде визначити сильні та слабкі сторони кожного окремого студента на основі того, як студент навчається онлайн, як і на які запитання зміг дати відповідь, профіль у соціальних мережах тощо. Це дозволить створити міцніші групи, які дозволять студентам засвоювати інформацію та покращить групові результати.

1.5.2 Створення масових індивидувальних програм

Всі ці дані допоможуть створити індивідуальну програму для кожного окремого студента. Великі дані дозволяють розробити індивідуальні програми в коледжах та університетах, навіть якщо кількість студентів близька до 10 000. Цього можна досягти за допомогою поєднання онлайнового та автономного навчання. Це дасть студентам можливість розвивати власну програму, відвідуючи ті заняття, в яких вони зацікавлені, працюючи у своєму темпі, маючи можливість консультуватися з викладачами. Створення масових індивідуальних програм для освіти - складне завдання. Завдяки алгоритмам стає можливим відслідковувати та оцінювати кожного студента окремо.

Ми вже бачимо, що це відбувається з масовими відкритими онлайн курсами (МВОК). Коли Ендрю Енг викладав курс з машинного навчання в Стенфордському університеті, його слухало 400 студентів. Коли в 2011 році курс був перероблений під МВОК, він привабив близько 100 000 студентів. Щоб зібрати таку кількість слухачів офлайн йому знадобилося б 250 років. 100 000 студентів, що беруть участь у класі, генерують велику кількість даних, які можуть відкрити невідомі до того закономірності. Працюючи одночасно для 100 000 студентів, також потрібні правильні інструменти для обробки, зберігання, аналізу та візуалізації всіх даних, що беруть участь у курсі. На даний момент ці МВОК все ще є масовими, але в майбутньому вони можуть перетворитися в масові індивідуальні курси.

Маючи 100 000 студентів, що беруть участь в МВОК, цу дасть університетам можливість знайти найкращих студентів з усього світу. На основі індивідуального поведінки студентів, курсів що вони прослухали, їх соціального профілю та того як вони знаходять інформацію в мережі можна знайти кращих студентів.

1.5.3 Покращення навчання в режимі реального часу

Коли студенти починають працювати самостійно, величезна кількість часу що витрачається на загальні теми, розраховані на студентів різного рівня, можуть бути охоплені онлайн. Професор може відстежувати всіх студентів у режимі реального часу і розповідати про більш цікаві та глибокі теми. Таким чином студенти не будуть прив’язані до середнього рівня слухачів курсу, а викладач зможе розповісти більше корисної інформації, краще розібрати складні теми предмету.

Стеження за навчанням студентів в режимі реального часу допоможе у вдосконаленні електронних методичок та навчальних курсів. За допомогою алгоритмів можна буде визначити які розділи важко зрозуміти, які розділи легко і які розділи незрозумілі. На підставі того скільки часу потрібно для того щоб прочитати текст, скільки питань задають по цій темі, скільки посилань натискаються для отримання додаткової інформації тощо. Якщо ця інформація надана в режимі реального часу, автори можуть змінювати підручники для задоволення потреб студентів, тим самим поліпшуючи загальні результати.

Більш того, BD може дати уявлення про те, як кожен учень навчається індивідуально. Кожен учень навчається по-різному, і спосіб, у який навчається студент, впливає на підсумковий бал. Деякі студенти навчаються дуже ефективно, тоді як інші можуть бути надзвичайно неефективними. Коли матеріали курсу доступні в інтернеті, можна стежити за процесом навчання. Ця інформація буде використана для

створення індивідуальної програми для студента або надання зворотнього зв'язку в реальному часі і таким чином покращити їх результати.

1.5.4 Зменшення кількості відсівів, покращення результатів

Весь цей аналіз покращить результати учнів, а також, можливо, знизять рівень відмов у вищих навчальних закладах. Відсів дорого коштує як для навчальних закладів, так і для суспільства. Коли відбувається постійний моніторинг успішності студентів отримується миттєвий відгук. Це може допомогти зменшити показники відсіву.

Використання передбачувальної аналітики щодо всіх зібраних даних може дати освітньому інституту розуміння майбутніх результатів студентів. Ці прогнози можуть бути використані для зміни програми, якщо вона прогнозує погані результати для неї або навіть запускає аналіз сценарію навчання за запропонованою програмою, перш ніж вона буде запущена. Університети та коледжі стануть більш ефективними при розробці програми, яка покращить результати, а також мінімізує кількість невдалих спроб та помилок при впровадженні програми.

Після закінчення навчання можна продовжити моніторити студентів, щоб побачити, як вони показують себе на ринку праці. Коли ця інформація буде оприлюднена, це допоможе майбутнім студентам у виборі університету, факультету та спеціальністю.

Великі дані будуть революціонізувати навчальну галузь в найближчі роки. Все більше і більше університетів та коледжів звертаються до BD, щоб поліпшити загальні результати студентів.

Висновки до розділу

BD допомагає великим компаніям, країнам та суспільству в багатьох галузях - від виробництва до медицини і навчання. Вона допомагає економити гроші та приймати кращі рішення базуючись на інформації що не була доступна до цього.

РОЗДІЛ 2 ІСНУЮЧІ ПІДХОДИ

Аналіз настрою можна вважати класифікаційним процесом, як показано на рисунку 2.1. Існують три основні рівні класифікації: рівень документа, рівень речень та рівень аспектів. Рівень документу має на меті класифікувати документ на основі висновку про позитивний чи негативний настрій до нього. Документ вважається основною інформаційною одиницею (говорять про одну тему). Ціль визначення рівня речення - це класифікація почуттів, виражених у кожному реченні. Перший крок - визначити, чи є думка суб'єктивною чи об'єктивною. Якщо думка є суб'єктивною, рівень речення визначає, чи речення має позитивне чи негативне забарвлення. Вільсон та ін. [5] у своїй роботі зазначили, що настрій не обов'язково суб'єктивний за своїм характером. Проте, немає принципової різниці між класифікаціями рівня документа та речення, оскільки речення - це лише короткі документи. Проте класифікація тексту на рівні документа або на рівні речення не надає всіх необхідних деталей про думку щодо всіх аспектів суб'єкта, які необхідні в багатьох прикладних випадках. Для отримання цих даних треба перейти на рівень аспектів. Рівень аспектів прагне класифікувати почуття щодо конкретних аспектів об'єктів. Перший крок полягає у визначенні сутностей та їх аспектів. В одному реченні можуть висловлюватися різні думки щодо різних аспектів того самого суб'єкта, наприклад "Зовнішній вигляд цього ноутбуку не дуже привабливий, проте технічні характеристики набагато кращі ніж у конкурентів". У магістерській дисертації розглядаються всі три рівні.

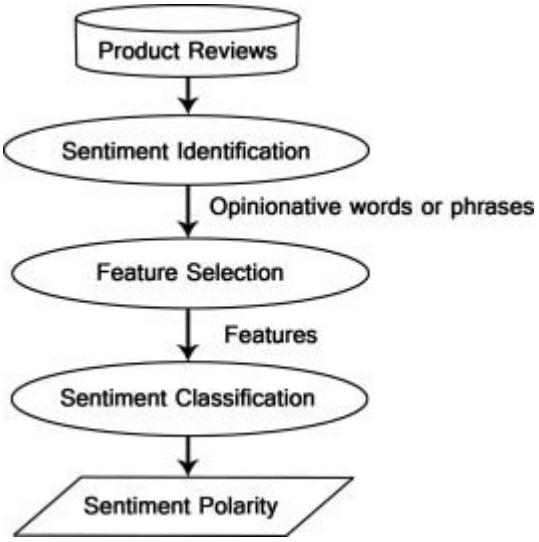

Рисунок 2.1 - Процес аналізу настроїв за відгуками продукту

Як дані будуть використані відгуки про продукти та сервіси. Вони важливі для бізнесу, оскільки можуть допомогти приймати бізнес-рішення відповідно до результатів аналізу думки користувачів про свою продукцію. Джерелами відгуків є, насамперед, відгуки з сайтів інтернет торгівлі (наприклад амазон або розетка), агрегатори відгуків (наприклад відгуки про вищі навчальні заклади). SA застосовується не тільки для відгуків про продукти, але й також на фондових ринках, новинах, або політичних дебатах. Наприклад, в політичних дебатах можна було б з'ясувати думки людей про певних кандидатів або політичних партій. Результати виборів також можна прогнозувати з політичних постів. Соціальні мережі та сайти мікро-блогів - гарні джерела інформації, оскільки люди обмінюються думками вільно, без жодної цензури.

За останні кілька років було запропоновано багато вдосконалень до алгоритмів SA. У цьому розділі ми ближче подивимося на ці вдосконалення, а також узагальнимо та систематизуємо деякі статті, представлені у цій галузі відповідно до різних методів SA. На рисунку 2.2 показана схема класифікації існуючих методів.

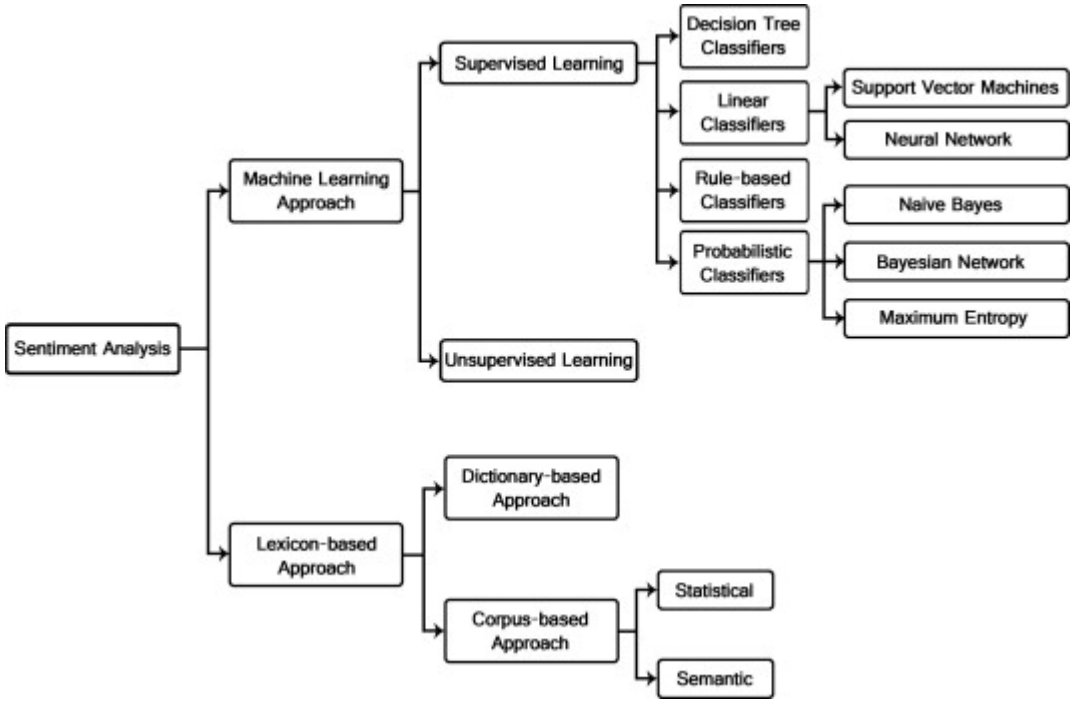

Рисунок 2.2 - Методи класифікації настрою.

Завдання аналізу настроїв розглядається як проблема класифікації почуттів. Першим кроком у проблемі визначення настрою є витяг та вибір текстових характеристик.

Характеристики:

- Частота та присутність термів: ці характеристики - це окремі слова або слова n грами та частота їх появи. Характеристика може мати різне числове значення: 1 якщо терм присутній в тексті та 0 якщо ні, або можна використовувати частоту появи терму в тексті, щоб вказати відносну важливість характеристики.
- Частини мови (ЧМ): пошук прикметників, оскільки вони є важливими показниками думок.
- Слова та фрази: це слова, які часто використовуються для висловлення думок, включаючи добрі або погані, наприклад, ненависть.
- Заперечення: поява заперечення може змінити орієнтацію настрою(наприклад ‘не добре’ рівнозначно ‘погано’).

2.1 Методи вибору функції

Методи відбору характеристик можна розділити на лексичні методи, що потребують втручання людини, а також статистичні методи, які є автоматичними методами, які застосовуються частіше. Лексичні підходи зазвичай починаються з невеликого набору базових слів. Потім цей набір завантажується та проводиться виявлення синонімів щоб розширити його. Цей напрямок має свої недоліки. З іншого боку, статистичні підходи є повністю автоматичними.

Методи вибору характеристик обробляють документи як групу слів – BOW – або як рядок, що зберігає послідовність слів у документі. BOW використовується найчастіше оскільки він простотіший в класифікації. Найпоширенішим кроком вибору характеристик видалення стоп–слів і приведення слова до свого кореня (наприклад сіл → село).

2.1.1 Точкова взаємна інформація

Міра взаємної інформації це формальний спосіб моделювання взаємної інформації між характеристиками та класами. Ця міра отримана з

теорії інформації. PMI $\square_{\square}(\square)$ між словом $\square$ та класом $\square$ визначається на основі рівня співпадіння між класом $\square$ і словом $\square$. Очікувана одночасна поява класу $\square$ і слова $\square$ на основі взаємної незалежності отримується наступним чином: $\square_{\square} \cdot \square(\square)$, а істинна одночасна поява - $\square(\square) \cdot \square_{\square}(\square)$.

Взаємна інформація визначається на основі співвідношенні між цими двома значеннями і задається наступним рівнянням:

$$\square_{\square}(\square) = \square\square\square\left(\frac{\square(\square) \cdot \square_{\square}(\square)}{\square(\square) \cdot \square_{\square}}\right) = \square\square\square\left(\frac{\square_{\square}(\square)}{\square_{\square}}\right), \quad (2.1)$$

де $\square(\square)$ ймовірність появи слова $\square$.

Слово $\square$ позитивно корелює з класом $\square$, коли $\square_{\square}(\square)$ більше 0. Слово $\square$ негативно корелює з класом $\square$, коли $\square_{\square}(\square)$ менше 0.

PMI використовується у багатьох програмах, і до нього застосовуються деякі вдосконалення. PMI розглядає лише спільну появу. Ю і Ву розширили основні PMI шляхом розробки контекстної ентропії, щоб розширити набір базових слів, створених з невеликого набору статтей про біржові новини. Їх модель контекстуальної ентропії вимірює подібність між двома словами, порівнюючи їх контекстні розподіли за допомогою міри ентропії, що дозволяє виявляти слова, подібні до базових. Після того, як базові слова були розширені, як базові, так і розширені слова використовуються для класифікації настроїв статей. Їх результати показали, що метод може виявити більш корисні емоційні слова, а їх відповідна інтенсивність покращує класифікацію. Запропонований метод показав кращі результати ніж методи PMI на основі методів розширення, оскільки вони враховують як частоту спільної появи, так і контекстний розподіл, отримуючи корисніші емоційні слова та менше слів-шуму.

2.1.2 Хі-квадрат

Нехай □ - загальна кількість документів у збірці, □□(□) - умовна ймовірність класу □ для документів, які містять □, □□ - глобальна частка документів, що містять клас □, і □(□) - глобальна частка документів, що містять слово □. Тому χ^2-статистика слова між словом □ та класом і визначається як:

$$\chi^2_{□} = \frac{□ \cdot □(□)^2 \cdot (□□(□) - □□)^2}{□(□) \cdot (1 - □(□)) \cdot □□ \cdot (1 - □□)}, \qquad (2.2)$$

χ^2 та PMI - два різні способи вимірювання кореляції між термінами та категоріями. χ^2 краще, ніж PMI, оскільки це нормоване значення; отже, ці значення можна використовувати при порівнянні в рамках однієї категорії.

χ^2 використовується у багатьох прикладних рішеннях; одна з них - це контекстна реклама. Вони виявили інтереси блогерів, щоб покращити контекстну рекламу. У дослідженні використовували реальну контекстеу рекламу з сайтів ebay.com, wikipedia.com та epinions.com. Вони використовували SVM для класифікації та χ^2 для FS. Їхні результати показали, що розроблений метод може ефективно ідентифікувати ті реклами, які позитивно корелюються з особистими інтересами блогера.

Хагенау і Лібман використовували функції зворотного зв'язку, використовуючи відгуки на ринку як частину процесу вибору характеристик для даних фондового ринку. Потім вони використовували їх з χ^2 та BNS. Вони показали, що надійний вибір характеристик дозволяє значно підвищувати класифікаційну точність при поєднанні із складними

типами характеристик. Їхній підхід дозволяє вибирати семантично важливі ознаки і зменшує проблему перенавченості при застосуванні підходу до машинного навчання. Вони використовували SVM як класифікатор. Їх результати показали, що поєднання розширених методів визначення характеристик та їх вибір на основі зворотного зв'язку підвищує точність класифікації та дозволяє поліпшити аналізи настроїв. Це пояснюється тим, що їхній підхід дозволяє скоротити кількість менш інформативних характеристик, тобто шуму, і обмежує негативні наслідки надмірного наближення при застосуванні машинного навчання для класифікації текстових повідомлень.

2.1.3 Латентне семантичне індексування (LSI)

Методи виділення характеристик намагаються зменшити розмірність даних, вибираючи лише вагомі з вихідного набору атрибутів. Методи перетворення значень створюють менший набір характеристик як функцію вихідного набору характеристик. LSI є одним із відомих методів трансформації функцій. Він переводить текстовий простір на нову систему координат, яка є лінійною комбінацією вихідних характеристик слова. Для досягнення цієї мети використовуються основні методи аналізу компонентів (PCA). Він визначає систему координат, яка зберігає найбільший рівень інформації про варіації значень основних характеристик. Основним недоліком LSI є те, що це безконтрольна техніка, і вона сліпа по відношенню до основного розподілу класів. Отже, характеристики, знайдені в LSI, не обов'язково є тими напрямками, за

якими можна досягти кращого розділення класового розподілу документів.

2.2. Основні виклики у визначенні характеристик

Дуже складним завданням зі знаходження характеристик є визначення іронії. Мета - визначення іронічних відгуків. Ця робота була запропонована Рейесом і Россо [7]. Вони мали на меті визначити характеристичну модель, щоб відобразити частину суб'єктивних знань, що лежить в основі таких оглядів та спроб описати виразні характеристики іронії. Вони створили модель, що представляє словесну іронію з точки зору шести категорій особливостей: n-грам, POS-грам, позитивний/негативний профіль та профілювання приємності. Вони побудували доступний для всіх наборів даних з іронічними відгуками з новинних статей, сатиричних статей та оглядів клієнтів, зібраних з сайту amazon.com. Вони були розміщені на основі вірусного ефекту в Інтернеті, тобто того вмісту, який викликає ланцюгову реакцію у людей. Вони використовували NB, SVM та DT для цілей класифікації. Їхні результати з трьома класифікаторами є задовільними, як з точки зору точності, так і відгуку та F-міри.

2.2 Методи класифікації настрою

Методи класифікації настрою можна грубо розділити на підхід до машинного навчання, лексичний підхід та гібридний підхід. Підхід до

машинного навчання застосовує відомі алгоритми ML та використовує лінгвістичні функції. Підхід, що ґрунтується на лексиці, спирається на лексикон настрою, колекції відомих і попередньо підготованих виразів настрою. Він поділяється на словниковий підхід та підхід, оснований на ядрі, який використовує статистичні або семантичні методи для визначення полярності почуття. Гібридний підхід поєднує в собі обидва підходи і дуже поширений з поглядами на лексикони, які відіграють ключову роль у більшості методів.

Методи класифікації тексту з використанням методу ML можна грубо розділити на методи, що підлягають нагляду та без нагляду. Методи з наглядом використовує велику кількість маркованих навчальних документів. Безконтрольні методи використовуються, коли класифіковані навчальні документи відсутні або наявні в невеликому обсязі.

Лексиконізований підхід залежить від знаходження лексикону відношення, який використовується для аналізу тексту. У цьому підході існують два методи. Словниковий підхід, який залежить від пошуку базису слів за допомогою яких висловлюють думку, а потім шукає словник їх синонімів та антонімів. Підхід, що базується на ядрі, починається з переліку посилань на думку, а потім знаходить інші слова думки у великому ядрі, щоб допомогти знайти слова думки з контекстними орієнтаціями. Це можна зробити, використовуючи статистичні або семантичні методи. Нижче наведено коротке пояснення алгоритмів обох підходів.

2.2.1 Підхід з машинним навчанням

Підхід з використанням машинного навчання базується на використанні відомих алгоритмів задля вирішення задач SA як класифікація звичайного тексту що використовує синтаксичні та/або лінгівстичні особливості.

Визначення проблеми класифікації тексту: маємо множину тренувальних наборів □ = {□1, □2,…, □□}де кожному набору присвоєний клас. Класифікаційна модель пов'язана з функціями базової запису до одного з міток класу. Тоді для даного екземпляра невідомого класу, модель використовується для прогнозування мітки класу для нього. Проблема жорсткої класифікації полягає в тому, що до екземпляру призначається лише одна мітка. Проблема класичної класифікації полягає в тому, що екземпляру присвоюється ймовірнісне значення міток.

2.2.1.1 Підхід з наглядом

Підхід з наглядом залежить від наявності помічених навчальних документів. В літературі існує багато видів контролюючих класифікаторів. У наступних підрозділах ми коротко опишемо деякі найбільш часто використовувані класифікатори в SA.

Імовірнісні класифікатори використовують моделі суміші для класифікації. Модель суміші передбачає, що кожен клас є компонентом суміші. Кожна компонента суміші є генеративною моделлю, яка

забезпечує вірогідність відбору певного терміну для цього компонента. Такі класифікатори також називають генеративними класифікаторами. Три з найвідоміших імовірнісних класифікаторів обговорюються в наступних підрозділах.

2.2.1.2. Наївний Байєсівський класифікатор

Наївний Байєсівський класифікатор - найпростіший і найпоширеніший класифікатор. Наївна класифікаційна модель Байєса обчислює постеріорну вірогідність класу на основі розподілу слів у документі. Модель працює з визначенням характеристик BOW, яка ігнорує положення слова в документі. Він використовує теорему Байєса для прогнозування ймовірності того, що заданий набір характеристик належить до певного класу.

$$□(□□□□□|□□□□□□□□) = \frac{□(□□□□□)*□(□□□□□□□□|□□□□□)}{□(□□□□□□□□)}, \quad (2.3)$$

де $□(□□□□□)$ - попередня вірогідність позначки або ймовірність того, що випадкова функція встановить мітку. $□(□□□□□□□□|□□□□□)$ - це попередня вірогідність того, що даний набір характеристик класифікується як мітка. $□(□□□□□□□□)$ - попередня імовірність того, що заданий набір характеристик відбувся. Зважаючи на наївне припущення, яке стверджує, що всі функції є незалежними, це рівняння можна переписати таким чином:

$$P(c \mid x) = \frac{P(c) * P(x_1 \mid c) * P(x_2 \mid c) * \ldots * P(x_n \mid c)}{P(x)}, \quad (2.4)$$

Покращений класифікатор NB був запропонований Кангом та Юо для вирішення проблеми того, що точність позитивної класифікації до 10% вище, ніж точність негативної класифікації. Через це зменшується середня точність, коли точність двох класів виражається як середнє значення. Вони показали, що використання цього алгоритму з відгуками ресторанів звужує розрив між позитивною точністю та негативною точністю порівняно з NB та SVM.

2.2.1.4 Байєсівська мережа

Основним припущенням класифікатора BN є незалежність характеристик. Інше крайнє припущення - всі харктеристики взаємозалежні. Це призводить до моделі Байєсівської мережі, яка є спрямованим ациклічним графіком, вузли якого являють собою випадкові величини, а ребра є умовними залежностями. BN вважається повною моделлю для змінних та їх взаємозв'язків. Тому модель для повного спільного розподілу ймовірностей по всіх змінних визначена. У видобутку тексту обчислювальна складність BN дуже дорога; тому використовується не часто.

BN використовував Ернандес та Родрігес [8] для розгляду реальної проблеми, в якій відношення автора характеризується трьома різними (але пов'язаними) цільовими змінними. Вони запропонували використання

багатовимірних класифікаторів мережі Байєса. Він об'єднав різні цільові змінні в одній задачі класифікації, щоб використовувати потенційні відносини між ними. Вони розширили багатовимірні системи класифікації до напівконтрольованого домену, щоб скористатися величезною кількістю непозначеної інформації, доступної в цьому контексті. Вони показали, що їх напівконтрольований багатомірний підхід перевершує найпоширеніші підходи SA і що їх класифікатор є найкращим рішенням у підпорядкованій структурі, оскільки він відповідає реальній структурі домену.

2.2.1.5 Класифікатор максимальной ентропій

Класифікатор максимальної ентропії (ME) (відомий як умовний експонентний класифікатор) перетворює позначені набори характеристик у вектори за допомогою кодування. Цей кодований вектор потім використовується для обчислення ваг для кожної характрестики, які потім можна об'єднати, щоб визначити найімовірніший клас для набору характеристик. Цей класифікатор параметризується набором □{□□□□□□□}, який використовується для об'єднання спільних характеристик, які генеруються з функції, встановленого кодуванням □{□□□□□□□□}. Зокрема, кодування відображає кожну □({□□□□□□□□□□,□□□□□})пару до вектора. Потім імовірність кожного класу обчислюється за допомогою наступного рівняння:

□(□□ | □□□□□) = (2.4)

□□□□□□□(□□□□□□□,□□□□□□(□□,□□□□□))
□□□(□□□□□□□(□□□□□□□,□□□□□□(□□,□)),□□□□□□□□□□□□)’

Цей класифікатор був використаний Кауфманом для виявлення паралельних речень між будь-якими мовними парами з невеликою кількістю навчальних даних. Інші інструменти, розроблені для автоматичного вилучення паралельних даних з непаралельних, використовують специфічні для мови техніки або вимагають великої кількості навчальних даних. Їхні результати показали, що класифікатори ME можуть давати корисні результати практично для будь-якої мовної пари.

2.2.1.6 Лінійні класифікатори

Маємо $\underline{□} = \{□_1, \ldots, □_□\}$ - це нормалізована частота появи слів документа, вектор $\underline{A} = \{a_1, \ldots, a_□\}$ - це вектор лінійних коефіцієнтів з тією ж розмірністю, що і функціональний простір, а □ - скалярний; висновок лінійного предиктора визначається як $□ = \underline{□} \cdot \underline{□} + □$, що є виходом лінійного класифікатора. Предиктор p - це роздільча гіперплоща між різними класами. Є багато видів лінійних класифікаторів; серед них - SVM, яка є формою класифікаторів, які намагаються визначити хороші лінійні розділювачі між різними класами.

2.2.1.7 Підтримка векторних класифікаторів машин (SVM)

Основним принципом SVM є визначення лінійних розділювачів у пошуковому просторі, які найкраще відокремлюють різні класи. На рисунку 2.3 є 2 класи □, □, і є 3 гіперплощини A, B і C. Гіперплощина A забезпечує найкраще розподіл між класами, тому що нормальна відстань будь-якої точки даних є найбільшою.

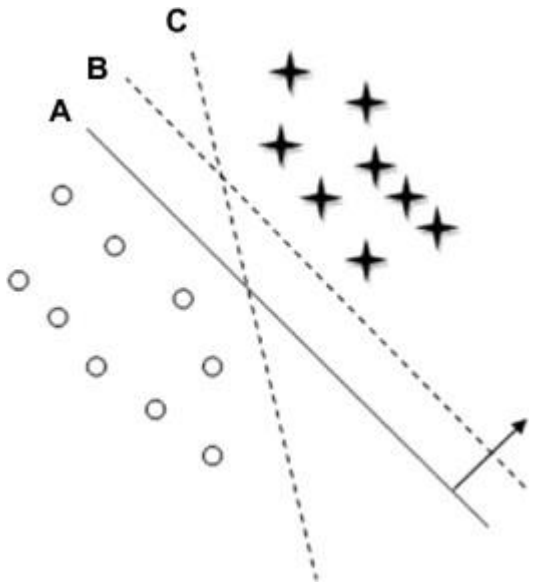

Рисунок 2.3 - Використання векторної машини підтримки на задачу класифікації

Текстові дані ідеально підходять для класифікації SVM через рідкісну властивість тексту, в якому деякі характеристики не мають відношення, але вони, як правило, корелюються один з одним і, як правило, організовані в категорії, які можна розділити лінійно. SVM може побудувати нелінійну поверхню рішення у початковому характеристичному просторі шляхом відображення примірників даних нелінійно до внутрішнього простору продукту, де класи можна розділити лінійно гіперплощиною.

SVMs були використані Лі і Лі як класифікатор настроїв. На відміну від бінарної класифікації, вони стверджували, що також слід враховувати суб'єктивність думки та експресивну довіру. Була запропонована структура, яка забезпечує компактний числовий підсумок думок на платформах мікро-блогів. Вони визначили та вилучили теми, згадані в думках, пов'язаних із запитами користувачів, а потім класифікували думки, використовуючи SVM. Вони працювали над публікаціями в twitter для свого експерименту. Результати показали, що розгляд довіри користувачів та суб'єктивності думок є важливим для агрегування думок мікроблогів. Вони довели, що їхній механізм може ефективно відкрити ринковий інтелект для підтримки осіб, що приймають рішення, шляхом створення системи моніторингу для відстеження зовнішніх поглядів на різні аспекти бізнесу в режимі реального часу.

2.2.1.8 Нейронна мережа

Нейронні мережі складаються з багатьох нейронів, і нейрон є її основною одиницею. Входи до нейронів позначаються вектором $\underline{\square_{\square}}$, який є частотою слів у і-му документі. Існує набір ваг A, які асоціюються з кожним нейроном, який використовується для обчислення функції його входів $\square(\cdot)$. Лінійна функція нейронної мережі: $\square_{\square} = \square \cdot \underline{\square_{\square}}$. У бінарній задачі класифікації передбачається, що позначення класу $\underline{\square_{\square}}$ позначається $\square_{\square}$, а знак передбачуваної функції $\square_{\square}$ призводить до присвоєння мітки цьому класу.

Багатошарові нейронні мережі використовуються для нелінійних меж. Ці множинні шари використовуються для індукування декількох

кусочно-лінійних меж, які використовуються для наближення замкнутих областей, що відносяться до певного класу. Виходи нейронів у попередніх шарах подаються в нейрони наступних. Процес навчання є більш складним, тому що помилки повинні повторно поширюватися на різних рівнях. Існує реалізація NN для текстових даних.

Існує емпіричне порівняння SVM та штучних нейронних мереж, представлених Моресом та Валяті, щодо аналізу настроїв на рівні документів. Вони зробили це порівняння, оскільки SVM широко і успішно використовується в SA, тоді як NN було приділено мало уваги як засобу для вивчення настроїв. Вони обговорили вимоги, результуючі моделі та контексти, в яких обидва підходи забезпечують найкращий рівень точності класифікації. Також було прийнято стандартний контекст оцінки з використанням популярних методів контролю для відбору характеристик та зважування в традиційній моделі BOW. Їхні експерименти показали, що NN дає чудові результати для SVM за винятком деяких несбалансованих контекстів даних. Вони провели тестування на трьох контрольних наборах даних на фільмах, GPS навігаторах, відгуках про камери та книги від amazon.com. Вони довели, що експерименти на наборі даних про фільми NN перевершили SVM зі статистично значущою різницею. Вони підтвердили певні потенційні обмеження обох моделей, які рідко обговорювалися в літературі SA, такі як обчислювальні витрати SVM під час роботи та NN під час тренувань. Вони довели, що використання інформаційного посилення (дешевий в обчисленні метод вибору характеристики) може зменшити кількість обчислень як для мережі, так і SVM без істотного впливу на отриману точність класифікації.

2.2.1.9 Дерева рішень

Дерева рішень забезпечує ієрархічне розкладання простору тренувальних даних, в якому для розділення даних використовується умова про значення атрибута. Умова або припущення - наявність або відсутність одного або декількох слів. Розподіл простору даних здійснюється рекурсивно, доки листові вузли не матимуть певних, наперед заданих мінімальних значень записів, які використовуються для цілей класифікації.

Існують й інші види предикатів, які залежать від подібності документів для кореляції наборів термінів, які можуть бути використані для подальшого розподілу документів. Різні види розщеплень - розбиття на окремі атрибути, які використовують присутність або відсутність окремих слів або фраз на певному вузлі дерева для виконання розбиття. Розподіл мульти-атрибутів на основі схожості використовує документи або часті кластери слів і подібність документів з цими кластерами слів для виконання розподілу.

Реалізація дерева рішень в текстовій класифікації, як правило, є невеликими варіаціями стандартних пакетів, такі як ID3 та C4.5. Лі і Джейн [9] використали алгоритм C5, який є наступником алгоритму C4.5. Залежно від поняття дерева; був запропонований підхід, щоб видобути структури змісту точних термінів у контексті рівня речення, використовуючи структуру MST, щоб виявити зв'язки між тематичним терміном ″□″та його контекстом. Відповідно, вони розробили так звану тематичну модель опису для класифікації почуттів. У їхньому визначенні "актуальні терміни" - це вказані об'єкти або певні аспекти об'єктів у певному домені. Вони запровадили автоматичне вилучення актуальних термінів з тексту на основі термінального терміналу домену. Далі ці

вилучені терміни використовувалися, щоб диференціювати теми документів. Ця структура передає інформацію про настрої. Їх підхід відрізняється від звичайних алгоритмів обробки дерев машинного навчання, але здатний ефективно вивчати позитивні та негативні контекстні знання.

Підхід що базується на теорії графів був представлений Яном і Бінгом [10]. Вони представили підхід до розповсюдження, щоб врахувати зовнішні та внутрішні до речення характеристики. Ці дві характеристики речення є внутрішньодокументними доказами та міждокументними доказами. Вони сказали, що для визначення орієнтації настрою необхідно більше інформації, ніж характеристики всередині самого речення. Вони працювали на домені камери та порівнювали свій метод як з безконтрольним підходом, так і з контрольованими підходами (NB, SVM). Їхні результати показали, що запропонований підхід краще, ніж обидва підходи, якщо вони не використовують зовнішніх характеристик висловлювання та перевершує попередні підходи.

2.2.1.10 Класифікатори на основі правил

У класифікаторах, заснованих на правилах, простір даних моделюється набором правил. Ліва сторона являє собою умову набору характеристик, виражену в диз'юнктивній нормальній формі, а праворуч - позначка класу. Умова - присутність певного терму. Відсутність терму рідко використовується, оскільки вона не інформативна у розріджених даних.

Існує декілька критеріїв для створення правил, етап навчання конструює всі правила залежно від цих критеріїв. Найбільш загальними

критеріями є підтримка та впевненість. Підтримка являє собою абсолютну кількість примірників у наборі навчальних даних, які мають відношення до правила. Впевненість стосується умовної ймовірності того, що права частина правила виконується, якщо ліва частина виконується.

Дерева рішень та правила прийняття рішень, як правило, переносять правила на простір характеристик, але дерево рішень прагне досягти цієї мети за допомогою ієрархічного підходу. Кіньлан [11] вивчив проблеми вирішення проблеми дерева рішень і вирішення проблеми в рамках єдиної системи; як певний шлях у дереві рішень можна вважати правилом класифікації текстового екземпляра. Основна відмінність між деревами рішень та правилами прийняття рішень полягає в тому, що DT являє собою строгий ієрархічний розподіл простору даних, тоді як класифікатори, що базуються на правилах, дозволяють перетини в просторі рішення.

2.3 Слабо, напів і безконтрольне навчання

Основною метою класифікації тексту є класифікація документів на певну кількість визначених категорій. Для цього використовується велике число позначених навчальних документів для керованого навчання. У текстовій класифікації іноді важко створювати ці помічені навчальні документи, але легко збирати не помічені документи. Безконтрольні методи навчання застосовуються для подолання цих труднощів. У цій галузі були представлені численні дослідницькі роботи, в тому числі робота Ко і Сео [12]. Вони запропонували метод, який розподіляє документи на речення, і категоризує кожне речення, використовуючи списки ключових слів кожної категорії та міри подібності речень.

2.3.1 Мета-класифікатори

У багатьох випадках дослідники використовують один чи більше класифікаторів для перевірки своєї роботи. Вони представили підхід ML для вирішення проблеми пошуку документів, що містять позитивну або негативну оцінку в медіа-аналізі. Дисбаланс у розподілі позитивних та негативних зразків, зміни документів у часі, а також ефективні методи навчання та оцінки для моделей - це завдання, з якими вони зіткнулися, щоб досягти своєї мети. Вони працювали над трьома наборами даних, створеними компанією з аналізу засобів масової інформації. Вони класифікували документи двома способами: виявлення присутності прихильності та оцінки негативної чи позитивної прихильності. Було використано п'ять різних типів характеристик, щоб створити набори даних з вихідного тексту. Вони перевірили багато класифікаторів, щоб знайти найкращий, який є (SVM, K-найближчий сусід, NB, BN, DT, правило, що вивчає та інші). Вони показали, що збалансування розподілу класів в навчальних даних може бути корисним для покращення продуктивності, але НБ може вплинути негативно.

Застосування алгоритмів ML для потокового передавання даних з Twitter було досліджено Руй і Лю. У своїй роботі вони досліджували, чи впливає "Твіттер із уст в уста" (WOM) на продажі фільмів, оцінюючи модель динамічної панелі даних. Вони використовували NB та SVM для цілей класифікації. Їхній основний внесок полягав у класифікації твітів з урахуванням унікальних характеристик твітів. Вони розрізняли попередню думку споживачів (тих, хто ще не купив товар), а також думку споживачів

(тих, хто купив товар). Вони зібрали дані Twitter WOM, використовуючи Twitter API та дані про продажі фільмів з BoxOfficeMojo.com. Їхні результати дозволяють припустити, що ефект WOM на продаж продуктів від користувачів Twitter із більшою кількістю послідовників значно перевищує вплив користувачів Twitter із меншим кількістю підписок. Вони виявили, що ефект попереднього споживання WOM на продаж фільмів більший, ніж пост-споживання WOM.

Інша стаття порівнювала багато класифікаторів після застосування класифікатора, основаного на статистичних марківських моделях. Вони були використані щоб врахувати залежності між словами та для створення словника, що покращило передбачувану ефективність кількох популярних класифікаторів. Це дослідження провів Баєм, який представив двоетапний алгоритм прогнозування. На першому етапі його класифікатор навчився умовним залежностям серед слів і кодував їх у Марківському ковдрі, спрямованому ациклічному графі для змінної настрою. На другому етапі він використовував мета-евристичну стратегію для точного налаштування алгоритму, щоб отримати більш високу перехресну перевірку точності. Він працював над двома колекціями інтернет-фільмів з IMDB та трьома колекціями онлайнових новин, потім порівняв його алгоритм із SVM, NB, ME та іншими. Він проілюстрував, що його метод вдалося виявити приблизний набір прогностичних характеристик й отримати кращі результати прогнозування настрою порівняно з іншими методами. Його результати свідчать про те, що почуття затьмарюються умовними залежностями між словами, а також за допомогою ключових слів або високочастотних слів. Складність його моделі лінійна від кількості зразків.

Підходи, що підлягають нагляду та без нагляду, можна об'єднати разом. Це зробили Вальдівіа та Камара [13]. Вони запропонували використовувати мета-класифікатори для розробки системи класифікації полярності настроїв. Вони працювали на наборі рецензій на фільми

іспанською разом з паралельним набором, перекладеним англійською. По-перше, вони створили дві індивідуальні моделі, що використовують ці два набори, а потім застосовують алгоритми машинного навчання (SVM, NB, C4.5 та інші). По-друге, вони інтегрували базис настроїв SentiWordNet в англійський набір, створюючи нову бездистанційну модель, використовуючи семантичний орієнтаційний підхід. По-третє, вони об'єднують три системи, використовуючи мета-класифікатор. Їхні результати перевершили результати використання окремих наборів і показали, що їх підхід можна вважати гарною стратегією класифікації полярності, коли доступні паралельні набори.

Класифікатори ML використовуються Уокер та Ананд [14] для класифікації позиції. Позиція визначається як загальна позиція, яку людина має до об'єкта, ідеї чи позиції. Позиція схожа на точку зору чи перспективу, її можна розглядати як ідентифікацію "сторони", на якій знаходиться персона, наприклад за чи проти певних політичних рішень. Уокер та Ананд [14] класифікують позицію, застосовуючи свій алгоритм до політичних дебатів. Вони використовували 104 двосторонні дебати від convinceme.net для 14 різних тем дебатів і намагалися визначити позицію або відношення ораторів. Їх головна мета - визначити потенційний внесок у визачення характеристик контекстного діалогу. Основний ефект для контексту полягає у порівнянні їх результатів з результатами що не враховують контекст, де лише 5 пар з характеристиками-темами показують зменшення при переході від безконтекстного аналізу до аналізу з контекстом. Вони використовували SVM, NB та класифікатор на основі правил. Вони досягли точності у визначенні сторони, для кожної теми, вище, ніж вихідні лінії уніграми при використанні настроїв, суб'єктивності, залежностей та діалогічних особливостей.

2.4 Лексичний підхід

Слова що висловлюють думки використовуються у багатьох завданнях класифікації почуттів. Слова позитивного сприйняття використовуються для вираження деяких бажаних станів, тоді як слова негативного сприйняття використовуються для вираження деяких небажаних станів. Існують також фрази і вирази думок, які разом називаються лексикою думок. Існує три основні підходи до складання або виправлення списку слів. Ручний підхід займає дуже багато часу, і він не використовується самостійно. Він, як правило, поєднується з двома іншими автоматизованими підходами як остаточна перевірка, щоб уникнути помилок, що виникають внаслідок автоматизованих методів. Розглянемо два автоматизованих підходи.

2.4.1 Словниковий підхід

Невеликий набір слів з лексики думок збирається вручну. Потім цей набір збільшують шляхом пошуку в добре відомих словниках WordNet або словниках за їх синонімами та антонімами. Нове знайдені слова додаються до базисного списку, після чого починається наступна ітерація. Ітераційний процес зупиняється, коли нових слів не знайдено. Після закінчення процесу можна провести ручну перевірку, щоб видалити або виправити помилки.

Підхід, оснований на словниках, має великий недолік, який полягає у нездатності знаходити лексикон думок з доменом та контекстними

орієнтаціями. Цю та Він [15] використовували підхід на основі словників для виявлення настрієвих речень в контекстній рекламі. Вони запропонували рекламну стратегію, спрямовану на покращення релевантності реклами та досвіду користувачів. Вони використовували словник синтаксичного аналізу та настрою і запропонували правило, засноване на вирішенні проблеми вилучення тем та ідентифікації споживачів при вилученні ключових слів із реклами. Вони працювали на веб-форумах з automotvieforums.com. Їх результати продемонстрували ефективність пропонованого підходу щодо вилучення ключових слів і вибору об'яв.

2.4.2 Корпусний підхід

Підхід, що ґрунтується на корпусі, допомагає вирішити проблему пошуку лексикону думок з контекстно залежними орієнтаціями. Методи залежать від синтаксичних моделей або моделей, які відбуваються разом із набором переліку базового лексикону думок, щоб знайти інші слова у великому корпусі. Один з цих методів починається з переліку прикметників базису та використовували їх разом з низкою лінгвістичних обмежень, щоб визначити додаткові слова прикметників думок та їхні орієнтації. Обмеження стосуються зв'язків, таких як І, АБО, АЛЕ, АБО-АБО тощо, наприклад І, говорить, що об’єднані прикметники зазвичай мають однакову орієнтацію. Ця ідея називається “послідовністю почуттів”, яка на практиці не завжди послідовна. Є також протилежні

вирази, такі як, “але”, “однак”, які позначаються як зміни у думці. Для того, щоб визначити, чи два прикметники з однаковими або різними орієнтаціями пов'язані, до великого корпусу застосовується навчання. Потім зв'язки між прикметниками утворюють граф, і на графі виконується кластеризація для створення двох наборів слів: позитивних і негативних.

Підхід на основі таксономії для вилучення думок функціональних рівнів та їх перетворення в функціональній систематиці був запропонований Крузом і Трояно [16]. Ця таксономія являє собою семантичне представлення частин та атрибутів об'єкта про які є враження. Їх головна мета - це орієнтована на домен ОМ. Вони визначили набір ресурсів, специфічних для домену, які містять цінні знання про те, як люди висловлюють думку щодо певного домену. Вони використовували ресурси, які автоматично були викликані з набору анотованих документів. Вони працювали на трьох різних областях (огляди навушників, готелів та автомобілів) від epinions.com. Їхні результати підтвердили важливість домену для побудови точних систем виявлення думок, оскільки вони призвели до підвищення точності щодо незалежних від доменів підходів.

Використання єдиного підходу, основаного на корпусі, не настільки ефективний, як підхід на основі словника, оскільки важко підготувати величезний корпус для охоплення всіх англійських слів, але цей підхід має важливу перевагу, яка може допомогти знайти доменні та контекстні думки та їх орієнтації за допомогою доменного корпусу. Підхід, що базується на корпусі, виконується за допомогою статистичного підходу або семантичного підходу, як показано в наступних підрозділах.

2.4.2.1 Статистичний підхід

Виявлення шаблонів спільної появи базисних слів можна зробити, використовуючи статистичні методи. Це може бути зроблено шляхом виведення постеріорної полярності за допомогою спільного виявлення прикметників у корпусі. Можна використовувати весь набір індексованих документів в Інтернеті як корпус для побудови словника. Це подолає проблему недоступності деяких слів, якщо корпус що використовується недостатньо великий.

Полярність слова може бути ідентифікована шляхом вивчення частоти виникнення слова у великому анотованому корпусі текстів. Якщо слово зустрічається частіше серед позитивних текстів, то його полярність позитивна. Якщо воно зустрічається частіше серед негативних текстів, то його полярність є негативною. Якщо воно має рівні частоти, то це нейтральне слово.

Подібні висловлювання часто з’являються разом у корпусі. Це основне зауваження, яке базується на сучасних методах. Тому, якщо два слова часто з’являються разом в одному контексті, вони, ймовірно, мають однакову полярність. Отже, полярність невідомого слова може бути визначена шляхом розрахунку відносної частоти співпадіння з іншим словом. Це можна зробити за допомогою PMI.

Статистичні методи використовуються у багатьох програмах, що стосуються SA. Один з них - виявлення маніпуляцій переглядів шляхом проведення статистичної перевірки випадковості, що називається тестовим пробігом. Ху і Боз [17] очікували, що стиль написання відгуків буде випадковим завдяки різним передумовам клієнтів, якщо відгуки були написані власне клієнтами. Вони працювали над оглядами книг від amazon.com і виявили, що близько 10,3% відгуків про продукти підлягають маніпуляціям.

Латентний семантичний аналіз (LSA) - це статистичний підхід, який використовується для аналізу взаємозв'язків між набором документів та

термінами, зазначеними в цих документах, для створення набору змістовних моделей, пов'язаних із документами та умовами. Цао і Дуань [18] використовували LSA для визначення семантичних характеристик з текстів огляду для вивчення впливу різних ознак. Мета роботи полягає в тому, щоб зрозуміти, чому деякі відгуки отримують багато корисних голосів, тоді як інші отримують мало. Тому замість прогнозування корисного рівня для відгуків, які не мають голосів, вони досліджували фактори, що визначають кількість голосів корисності, які отримують певний огляд (включають як «так», так і «ні»). Вони працювали над відгуками користувачів програмного забезпечення від CNET download.com. Було показано, що семантичні характеристики є більш впливовими, ніж інші характеристики, які впливають на те, скільки голосів корисних оцінок отримує відгук.

Семантична орієнтація слова - це статистичний підхід, який використовується разом із методом PMI. Існує також реалізація семантичного простору, що називається Hyperspace Analogue to Language (HAL), запропоноване Лундом і Берджессом [19]. **Семантичний простір** - це простір, в якому слова представлені точками; позиція кожної точки разом з кожною віссю якось пов'язана зі значенням слова. Підхід, заснований на HAL, який називається Sentiment Hyperspace Analogue to Language (S-HAL). У цій моделі змістовна орієнтація інформації про слова характеризується специфічним векторним простором, а потім класифікатор навчався виявляти семантичну спрямованість термінів (слів або фраз). Гіпотеза була підтверджена методом семантичної орієнтації виведення з PMI (SO-PMI). Їхній підхід створював набір зважених ознак, що базуються на оточуючих словах. Підхід був перевірений на коментарях до новин і використовували китайський корпус. Їхні результати показали, що вони перевершили SO-PMI і показали переваги при моделюванні

семантичних характеристик орієнтації в порівнянні з вихідною моделлю HAL.

2.4.2.2. Семантичний підхід

Семантичний підхід дає безпосереднє значення настрою та спирається на різні принципи обчислення подібності між словами. Цей принцип дає подібні почуття цінності семантично близьким словам. Наприклад, WordNet надає різні види семантичних зв'язків між словами, що використовуються для розрахунку полярностей настроїв. WordNet також може бути використаний для отримання списку сентиментальних слів, ітераційно розширюючи початковий набір з синонімами та антонімами, а потім визначаючи полярність почуття для невідомого слова за допомогою відносного підрахунку позитивних та негативних синонімів цього слова.

Семантичний підхід використовується в багатьох додатках для побудови моделі лексики для опису дієслів, іменників і прикметників, які будуть використовуватися в SA як робота, представлена Максом і Воссеном [20]. Їхня модель описує детальні суб'єктивності відносин між аспектами у реченні, що виражає окреме ставлення до кожного аспекту. Ці відносини суб'єктивності позначаються інформацією щодо ідентичності автора та орієнтації (позитивного чи негативного) ставлення. Їх модель включала категорізацію в семантичні категорії, що відносяться до SA. Це забезпечило засоби для ідентифікації власника ставлення, полярності ставлення, а також опису емоцій та настроїв різних суб'єктів, що беруть

участь у тексті. Вони використовували голландський WordNet у своїй роботі. Їхні результати показали, що суб'єктивність спікера, а іноді і суб'єктність автора, може бути визначена точно.

Семантичні методи можна змішувати з статистичними методами для виконання завдання SA як робота, представлена Чжаном та Сю [21], які використовували обидва способи для визначення слабкості продукту з огляду в Інтернеті. Їх шукач слабкості витягує характеристики та явні групові характеристики, використовуючи морфемний метод для ідентифікації функціональних слів з відгуків. Вони використовували західну подібність, щоб знайти часті та рідкісні явні функції, які описують один і той же аспект. Виділили неявні характеристики із методом вибору методу колаборації, який використовує PMI. Вони об'єднують продукти, вводячи слова в відповідні аспекти, застосовуючи семантичні методи. Була використана методику SA на основі речення для визначення полярності кожного аспекту у реченнях з урахуванням впливу ступеня прислівників. Таким чином метод допомагає визначити слабкі місця продукту.

Висновок до розділу

У даному розділі були розглянуті найпопулярніші методи визначення настрою автора та визначення аспектів. Були вказані переваги та недоліки кожного методу, задачі які вони вирішують.

РОЗДІЛ 3 АРХІТЕКТУРА ТА АНАЛІЗ РЕЗУЛЬТАТІВ РОБОТИ

У цьому розділі пропонується модифікований підхід, заснований на ідеях з аспектно-орієнтованого аналізу Бей Лю. Ідея модифікації полягає в тому, що лексика у відгуках відрізняється в залежності від продукту, про який пише автор. Оскільки підхід Лю направлений на огляди фізично існуючих продуктів (телевізори, телефони тощо), його не можна застосовувати до сфери послуг, в якій є особливості, що не були враховані в початковій моделі. Після детального ознайомлення з відгуками в сфері послуг (відгуки про вищі навчальні заклади), були знайдені особливості, і згодом додані в модель для розширення. У роботі запропоновано новіші та складніші NLP правила для класифікації суб'єктивних думок та настроїв щодо аспектів, що розглядаються. Також зроблена спроба вирішити проблему візуалізації результатів аналізу для допомоги кінцевим користувачам швидко та легко засвоювати результати аналізу. Робота включає розробку спільної архітектури для аспектно-орієнтованого аналізу. Результати показують, що запропонована модифікація здатна працювати краще, ніж модель Лю в сфері послуг, покращуючи точність класифікації суб'єктивності та настроїв. Підхід показав свою ефективність у визначенні настрою, досягаючи F-міри 92% для поставленої задачі. В середньому алгоритми були здатні визначити 35% з явних виразів думки щодо аспектів без використання розширення.

3.1 Введення

Модицікація використовує той факт, що під час написання відгуків в Інтернеті користувачі використовують різні слова для різних видів продукції. Візьмемо, наприклад, абстрактний продукт, який стосується концептуального товару, виробленого деякою індустрією. Як правило абстрактні продукти класифікують використовуючи дві категорії, фізичні товари та нематеріальні послуги. Більшість розглянутих у попередньому розділі робіт зосереджені лише на оглядах фізичних продуктів. У таких оглядах користувачі зазвичай одразу говорять про функції продукту, які їм сподобалися або не сподобалося. Проте для інших видів продукції все змінюється.

Роботи, вже обговорювали важливість домену в галузі виявлення думок. Коли людина пише огляди фільмів, вона, скоріше за все, коментує не лише елементи фільму, але й людей, пов'язаних із кіно. Після детального вивчення оглядів відгуків про вищі навчальні заклади, були знайдені особливості притаманні сфері освіти, які будуть використані в моделі запропонованої модифікації. Загалом користувачі, як правило, пишуть про свій досвід використання сервісу, коли залишають відгуки та використовують довші та складніші речення.

По-перше, багато речень містять більш ніж одну згадку про продукт, який розглядається, а також про його характеристики та складові. З іншого боку, багато речень не містять думок. Крім того згадуються об'єкти, які не відповідають характеристикам або компонентам розглянутого продукту. Ці речення, як правило, пояснюють досвід автора. Було помічено, що студенти можуть використовувати багато різних і складних виразів для позначення характеристик або частин продукту.

Ідея модифікації ґрунтується на ідеях Ланкастера [22], в яких сказано, що споживчі переваги продукту невід'ємно пов'язані з його

особливостями. Пропозиція полягає в тому, щоб дізнатись, які характеристики присутні, і визначити, яке ставлення автора стосовно цих характеристик. Це допоможе краще зрозуміти переваги, в термінах “подобається” та “не подобається”.

По-друге, в результаті аналізу домену були створені спеціальні набори даних, які допомагають відобразити його особливості. Також ці набори даних використовуються для оцінки запропонованих моделей для висновків, що базуються на аспекті. Крім цього робота також включала розробку загальної архітектури для інструменту виявлення думок щодо аспектів. Основна задача системи - допомогти користувачам зрозуміти ставлення та загальну оцінку сервісу, надати можливість легко знайти або отримати суб'єктивну інформацію з відгуків клієнтів, отриманих з відповідних сервісів або агрегаторів відгуків.

3.2 Довідкова інформація

Пояснимо моделі та ідеї Бін Лю, які лягли в основу запропонованого методу. Розглянемо підхід Лю в загальних рисах. На його думку, що висловлення характеризуються з 5 складових [23]:

- сутність - пропонується позначити об’єкт, про який йде мова, або іншими словами те, що оцінюється. Сутність може містити набір компонентів і атрибутів, і аналогічним чином, кожен компонент об'єкта може мати свої підкомпоненти та атрибути. Вона може бути розбита на дерево або ієрархію підатрибутів і підкомпонентів;
- аспект - оскільки важко вивчити сутність на рівні довільної ієрархії, ця ієрархія спрощена на один або два рівні, і атрибути або компоненти об’єкта позначаються як аспект. Таким чином, корінь

ієрархії або дерева - сутність, кожен лист є аспектом, і посилання є частиною відносин;

- орієнтація настрою - визначення того чи думка висловлена по відношенню до об’єкта позитивна або негативна;
- власник думки - користувач, що висловив думку;
- час - час та дата висловлення думки.

Таким чином, думки вважаються позитивним чи негативним поглядом, ставленням, емоціями або оцінкою щодо суб'єкта або аспекту цього суб'єкта від власника думки в певний час. Введено також наступні поняття:

- вираз сутності: відповідає дійсному слову або фразі, написаній користувачем для позначення або позначення сутності. У результаті об'єкти є узагальненням кожного виразу сутності, використаного в аналізованих документах, або конкретної реалізації виразу сутності;
- вираз аспекту: як і для виразу суб'єкта, аспектним виразом є фактичне слово або фраза, написані користувачем для позначення аспекту. Таким чином, аспекти також є загальними поняттями, які містять кожне вираження аспекту.

Тоді можна визначити модель суб'єкта та модель документа думки. Сутність $□_□$ відображається сам по собі як ціле а також як скінченний набір аспектів, $□_□ = \{□_{□1}, □_{□2}, \ldots, □_{□□}\}$. Сутність може бути виражена будь-яким з кінцевих наборів виразів сутності: $□□_□ = \{□□_{□1}, □□_{□2}, \ldots, □□_{□□}\}$. Кожен аспект $□_{□□}$ $□_□$ сутності може бути виражений будь-яким з кінцевих наборів виразів аспекту $□□_{□□} = \{□□_{□□1}, □□_{□□2}, \ldots, □□_{□□□}\}$. З іншого боку, документ думки $□_□ \in □$ містить думки про сукупність сутностей $□_1, □_2, \ldots, □_□$ з набору власників думок $□_1, □_2, \ldots, □_□$. Погляди на кожну сутність $□_□$ виражаються на самій сутності а також на підмножині її аспуктів $□_{□□}$.

3.2.1 Ідентифікація аспекту

Цей етап спрямований на пошук та визначення важливих тем у тексті, які потім будуть використані для узагальнення. У [24] Ху і Лі представляють техніку, засновану на NLP та статистиці. У їх пропозиціях частини мовлення (POS) і дерево синтаксичного аналізу використовуються для пошуку іменників й іменникових фраз (ІФ). Потім, базуючись на частоті, визначаються іменники та іменникові фрази що використовувалися найчастіше. Отримані набори імен та ІФ фільтруються за допомогою спеціальних лінгвістичних правил. Ці правила гарантують, що терміни в межах тих аспектів, які складаються з більш ніж одного слова, ймовірно, представляють собою опис одного об'єкта, а також усувають зайві аспекти. Вони також виводять нечастотні аспекти, використовуючи підхід, знаходячи іменники або ІФ, які з'являються поруч з думками з високою частотою. Цей підхід не визначає прикметників або будь-який інший вид аспектів що не стосуються об’єкту.

3.2.2 Прогноз почуття

Наступним етапом є прогнозування настроїв, визначення орієнтації настроїв по відношенню до кожного аспекту. Дін, Лю та Ю пропонують лексику та підхід, заснований на правилах [25]. Цей метод спирається на словник сентиментальних слів, який містить список позитивних та негативних слів, які використовуються для співставлення термінів у тексті. Крім того, оскільки інші спеціальні слова також можуть змінити

орієнтацію, пропонуються спеціальні мовні правила. Прикладами таких слів є заперечення слово "не", а також деякі загальні заперечення. Однак, незважаючи на те, наскільки прості ці правила здаються на перший погляд, дуже важливо використовувати їх з обережністю, оскільки не всі випадки використання таких правил матимуть однакове значення.

У цьому контексті правила, розроблені Дінгом, Лю і Ю, включають функцію агрегування, щоб визначити орієнтацію по відношенню до аспекту в реченні, що поєднує декілька слів для висловлення думок.

Така функція буде використана в запропонованому методі.

3.2.3 Генерація висновку

Останній крок - генерація висновку, щоб легко представити оброблені результати. Лю визначає своєрідне резюме, що називається підсумком думки на основі аспектів, що складається з базових графіків, що показують кількість позитивних та негативних оцінок щодо кожного аспекту однієї сутності. Стрічкові діаграми могли використовуватися для порівняння набору виділених продуктів, відображаючи набір усіх аспектів вибраних продуктів на графіку. У цьому випадку кожна стрічка, розташована над або під віссю X, може відображатися у двох масштабах:

- фактична кількість позитивних чи негативних оцінок, нормованих за максимальною кількістю думок щодо будь-якої характеристики будь-якого товару;
- відсоток позитивних чи негативних оцінок, порівнюючи аспекти з точки зору відсотка позитивних та негативних відгуків.

3.3 Запропонована модифікація

3.3.1 Витяг формату вираження

Аспекти не з'являються в тексті безпосередньо, але вони існують у формі виразів аспектів. Відповідно, коли намагається застосувати модель Лю до визначення думок на реальних даних, поняття, що використовуються, можуть бути дещо заплутаними або незрозумілими. Також незрозуміло, як обробити аспекти, що з'являються в документі декілька разів. Саме для вирішення цих питань була розроблена модель побудови для визначення думок в документі.

Щоб зробити все простіше, розглянемо набір документів, які мають певну полярність: $□_□ = \{□_{□1}, □_{□2}, \ldots, □_{□□}\}$ щодо однієї сутності, $□_□$. Це припущення відповідає реальній ситуації, оскільки думки, як правило, доступні у вигляді відгуків про продукти в Інтернеті. Тоді кожен підтверджений документ відповідатиме відгуку чи думці, наданому власником $□_□$ у певний час $□_□$. Нехай $□_{□□}$ буде сукупністю всіх речень в $□_{□□}$, з $□_{□□} = \{□_{□□1}, □_{□□2}, \ldots, □_{□□□}\}$. Думки про $□_□$ в $□_{□□}$ буде виражатись на самому суб'єкті і на підмножині $□_{□□}$його аспектів. Аналогічно, кожен аспект $□_{□□}$ з'явиться на $□_{□□}$ як множина виразів аспектів$□□_{□□□}$, підмножина $□□_{□□}$. Суб'єкт не буде відображатися як підмножина різних виразів сутності $□□_{□□} \subseteq □□_□$. Таким чином, встановлений $□□_{□□}$ визначається як сукупність усіх аспектних виразів усіх аспектів і всіх виразів сутності, що з'являються в $□_□$. Речення пов'язане з одним вираженням аспекту або вираженням сутності лише тоді, коли воно з'являється у цьому реченні. Потім, для кожної пари

(□□, □) потрібно визначити орієнтацію настроїв, лише якщо на ньому з'явиться будь-яке вираження аспекту або вираз. Після визначення орієнтації почуттів просто необхідно додати □□, □□ та відповідний документ □□□ до результату. З іншого боку, пропозиція Лю вказує на те, що часто використовувані іменники у відгуках на продукти, як правило, є справжніми і важливими виразами аспектів, тому що коли люди коментують різні аспекти продукту, словники, які вони використовують, зазвичай збігаються. Тим не менш існують дві основні причини, які підтверджують той факт, що багато різних виразів можуть мати одне й те саме значення:

- принцип економії в мовах вказує на те, що вони намагаються багато чого сказати, використовуючи меншу кількість слів. Наприклад, речення "На кафедрі є хороший комп’ютерний клас" відповідає лексикалізації, де оригінальний вираз "Кафедра має хороше технічне забезпечення для класів програмування", скорочується відповідно до принципу економії;

- кожна мова пропонує системи, які організовують свої концепції, а також здійснюють спрощення. З цієї причини багато слів англійською (як і всіма іншими мовами) просто є гіпонімами визначеного гіперним. Фразеологізми це слово або фраза, якої семантичне поле включається в тому, що іншого слова, його hypernym. Для екземплярів, червоний, вермиліон, кармін та малиновий - це всі гіпоніми червоного кольору (гіперним), що, в свою чергу, є гіпонімом кольору.

На практиці, пошук аспектів, які оцінюються в сукупності документів, що висвітлюються, є справді складним завданням. Фактично, виявлення аспектних виразів з безлічі документів з думкою є зовсім іншим завданням, ніж визначити або знайти реальні аспекти в них, оскільки кількість можливих виразів, що з'являються в тексті, дійсно величезна.

Ще одне питання, виявлене у пропозиціях Лю, пов'язане з поняттями речення та відстані слова, які, хоч і широко використовуються, не є чітко визначеними. Незважаючи на поглиблений лінгвістичний аналіз, тут ми визначимо пропозицію як упорядкований набір елементів, включаючи слова та пунктуацію. Один елемент, який з'являється в двох різних реченнях, повинен розглядатися двічі, оскільки речення, де вони з'являються, відрізняються. Іншими словами, речення □ **відповідатиме набору унікальних кортежів (токен, позиція). Позиції можуть бути тільки в N ∪ {0}, а різниця між двома сусідніми компонентами повинна бути рівна 1. Поняття відстані слова між двома елементами речення** □ буде відповідати різниці позицій двох елементів у □.

$$WD(t_a, t_b) = |position(t_a) - position(t_b)| \; t_a, t_b \in S \qquad (3.1)$$

де □□(□□, □□) **(відстань слова) є просто абсолютною величиною різниці між числами в N ∪ {0}, то відстань слова** (□□, □□) є метрикою на множині □, оскільки вона задовольняє умовам невід'ємності, симетрії та нерівність трикутника. Зверніть увагу, що мінімальна відстань між двома елементами в □ = *1*, і ця рівність справджується лише між сусідніми елементами. Максимальна відстань відповідає |□| + *1*. Незважаючи на ці визначення та формалізацію, в цій роботі ми зосередилися на задачі визначення орієнтації настроїв на аспектному рівні, тому тут ми просто застосовуємо техніку, розроблену Ху і Лю в [24], для виявлення частих аспектів. Іншими словами, прагнучи зробити простіший аналіз, ми будемо розглядати вирази аспектів лише як

іменники або набори іменників, які ми називаємо явними виразами аспектів. Неявні або нечастих вирази аспектів визначатися не будуть.

3.3.2 Визначення орієнтації думок

Беручи роботу [25] як базу, був розроблений набір правил для визначення орієнтації речення, завжди враховуючи лексикон думок як основу.

3.3.2.1 Правила орієнтації слів

Перш за все необхідно визначити орієнтацію кожного слова в реченні. Для цього запропоновано алгоритм 1, зображений на рисунку 3.1. Алгоритм застосовує набір лінгвістичних правил, які будуть пояснені нижче.

Algorithm 1 Word Orientation

```
if word is in opinion_words then
    mark(word)
    orientation ← Apply Opinion Word Rule(marked_word)
else
    if word is in neutral_words then
        mark(word)
        orientation ← 0
    end if
end if
if word is near a too_word then
    orientation ← Apply Too Rules(orientation)
end if
if word is near a negation_word then
    orientation ← Apply Negation Rules(orientation)
end if
return orientation
```

Рисунок 3.1 - Псевдокод алгоритму визначення орієнтації слова

- Правила слова: слова позитивного сприйняття мають оцінку 1, що позначає нормалізовану позитивну орієнтацію, тоді як негативні - оцінку з протилежним знаком -1. Кожен іменник і прикметник у

кожному реченні, яке не є думкою, матиме внутрішній бал 0 і буде називатися нейтральним словом.

- Правила заперечення: слово або фраза заперечення зазвичай змінює думку, виражену у реченні. Отже, лексикон думок або нейтральні слова, на які впливають негативні наслідки, повинні бути розглянуті спеціально. Треба застосувати три правила: Запепечення Негативу → Позитив, Заперечення Позитиву → Негатив і Заперечення Нейтрального → Негатив. Заперечення слова та фрази включає в себе: "ні", "ні", "ніколи", "не", "не", "не можу", "не зробив", "не буде", "немає", "не слід" (навмисно з орфографічними помилками).
- Правило надмірності: речення, в яких з'являються слова "занадто", "надмірно" або "забагато", також обробляються спеціально. Коли слово з лексикону думок або нейтральне слово з'являється біля одного з названих термінів, вони також позначаються і його орієнтація завжди буде негативною (оцінка = -1).

3.3.2.2 Правила орієнтації аспекту

Пояснивши правила, які допомагають визначити орієнтацію кожного слова в реченні, пояснимо, яким чином всі отримані орієнтації будуть об'єднані для визначення остаточної орієнтації речення щодо певного аспекту. Ідея агрегації підсумовується в алгоритмі 2, зображеного на

рисунку 3.2, і розглядаються лише слова, позначені як слова думки чи нейтральні слова, які ми називаємо позначеними словами, оскільки вони є єдиними, які надають орієнтацію кожному реченню.

Algorithm 2 Opinion Orientation

```
1: if but_word is in sentence then
2:     orientation ← Opinion Orientation(aspect,marked_words,but_clause)
3:     if orientation ≠ 0 then
4:         return orientation
5:     else
6:         orientation ← Opinion Orientation(aspect,marked_words,not but_clause)
7:         if orientation ≠ 0 then
8:             return -1 × orientation
9:         else
10:            return 0
11:        end if
12:    end if
13: else
14:    for all aspect_position in aspect do
15:        for all aspect_word in aspect_position do
16:            for all word in marked_words do
17:                suborientation += Word Orientation(word) / WD(aspect_word,word)
18:            end for
19:            orientation += suborientation
20:        end for
21:        final_orientation += orientation
22:    end for
23:    if final_orientation > 0 then
24:        return 1
25:    else
26:        if final_orientation < 0 then
27:            return -1
28:        else
29:            return 0
30:        end if
31:    end if
32: end if
```

Рисунок 3.2 - Псевдокод алгоритму визначення орієнтації думки

- Правило агрегування аспектних слів: нехай □ буде реченням, що містить набір виразів аспектів □ = {□$_1$,…, □$_□$}, кожен з яких з'являється в □ лише один раз. Крім того, нехай □□$_□$ - це сукупність слів, які містять аспект □$_□$, де □□$_□$={□□$_{□1}$, □□$_{□2}$, …, □□$_{□□}$}. Кожен □□$_{□□}$ буде називатися словом аспекту й буде відповідати виразу аспекту □$_□$. Якщо відомі результати для кожної

думки та нейтрального слова в □, то оцінка для кожного □□□□ в □ задана наступною функцією агрегації:

$$score(aw_{ij}, s) = \sum_{ow_j \in s} \frac{score(ow_j)}{WD(ow_j, aw_{ij})} \quad (3.2)$$

де □□□- це думка чи нейтральне слово в □, □□(□□□, □□□□) - це відстань слова між аспектним словом □□□□ та слова думок □□□ у □. Рядок 17 реалізує цю формулу в алгоритмі 2. Ми беремо цю функцію з [25]; проте їх пропозиція не пояснювала того, як функція повинна застосовуватися до аспектних виразів, що складаються з більш ніж одного слова (який ми називаємо сполукою). Ми бачили, що у відгуках про навчання деякі аспектні вирази насправді складні. Наприклад, у реченні "Курс був застарий, але предмет був цікавим". Вираз аспекту, який повинен бути витягнутий алгоритмами Лю, - предмет. Проте пропозиція Лю не пояснює, яким чином орієнтація на цей аспект має бути отримана з речення. Щоб розглянути ці випадки, ми пропонуємо, щоб формула використовувалася для кожного слова в реченні, а не для кожного виразу аспекту. Ці орієнтації об'єднуються відповідно до наступного правила:

- Правило агрегування аспектів: для кожного складового аспекту виразу □□ в □ його орієнтація буде розрахована, беручи до уваги оцінки всіх слів, які його складають, □□□□ ∈ □□□, згідно з наступним рівнянням, яке реалізовано у рядку 19 алгоритму 2.

$$score(a_i, s) = \sum_{aw_{ij} \in AW_i} score(aw_{ij}, s) \quad (3.3)$$

- Правило агрегації позиції: ми також помітили, що у рецензіях на навчання аспектні вирази можуть з'являтися неодноразово в реченні. Цей випадок не покривається підходои Лю, але нам потрібен спосіб охопити такі ситуації. Припускаючи, що □□ з'являється □ разів у □ і знаючи оцінку виразу кожного виразу аспекту □□□, □ ∈ {*1,2,…,* □}, ми пропонуємо, щоб остаточна оцінка □□ або □□□□□□(□□, □) слід обчислити шляхом простого додавання значень балів усіх появ □□ в □ згідно з наступним рівнянням.

$$fscore(a_i, s) = \sum_{k=1}^{t} score(a_i^k, s) \qquad (3.4)$$

Формула з'являється у рядку 21 алгоритму 2. Варто зазначити, що коли □□ з'являється лише один раз у □, □□□□□□(□□, □) = □□□□□(□□, □). Рядки 23 - 31 показують, як орієнтація обчислюється відповідно до □□□□□ кожного аспектного виразу. Якщо □□□□□□(□□, □) є додатнім, думка щодо □□ вважається позитивною (рядки 23 та 24), і якщо він є від'ємним, думка вважається негативною (рядки 26 та 27). Інакше речення вважається нейтральним (рядок 29).

- Правила "але": ми використовуємо точно таке ж правило, запропоноване в [25]. Це правило означає, що коли зустрічається "А, але В" (включаючи будь-який синонім "але") з'являється у реченні □, □ повинно бути розбите на два сегменти, перший до "але", другий - після. Якщо орієнтація будь-якого аспектного слова □□□□,

що з'являється в сегменті речення після B дорівнює нуль, то його орієнтація повинна бути визначена з використанням сегмента до B, але з протилежним значенням. Ми зрозуміли, що існує невелика двозначність, оскільки в деяких випадках □□□□ може з'явитися поза розглянутим сегментом. У такому випадку наш підхід пропонує додати □□□□ до кінцевої позиції відповідного сегменту, щоб уникнути проблеми узгодженості. Рядки 1 - 12 в алгоритмі 2 застосовують це правило.

3.3.3 Підведення підсумків

Пропозиція Лю здається досить простим та ефективним підходом для узагальнення думок. Проте в ньому відсутній надійний спосіб вимірювати важливість кожного оціненого аспекту. Аспекти класифікуються відповідно до частоти їх появи в оглядах, але також можливі інші типи ранжування, наприклад ранжувати аспекти відповідно до кількості позитивних чи негативних оцінок. У запропонованому підході важливість кожного аспекту визначається одночасно, використовуючи кількість позитивних та негативних оцінок про нього. Ця міра також використовується для оцінки аспектів. Основне припущення полягає в тому, що аспект, який має багато позитивних та негативних оцінок, буде більш важливим, оскільки висока кількість думок обох орієнцій може означати, що користувачі дуже зацікавлені в цьому аспекті. Таким чином, загальна кількість разів, коли аспект з'являється, розглядається не тільки в оцінці важливості, а й у розрізненні кількості позитивних та негативних оцінок. Нехай □□і □□ - це кількість позитивних та негативних оцінок щодо аспекту □□, □ ∈ {*1*, … , □}. Тоді □□□□□□□ та □□□□□□□ будуть

нормалізованими значеннями □□ та □□, відповідно. При цьому обчислимо стандартне відхилення оцінок наступним чином:

$$AVScore_i = \frac{PScore_i + NScore_i}{2} \tag{3.5}$$

$$STDScore_i = \sqrt{\frac{(PScore_i - AvScore_i)^2 + (NScore_i - AvScore_i)^2)}{2}} \tag{3.6}$$

Нова міра для кожного виразу аспекту □□, яку ми назвали як “відносна важливість”, визначається як мінімальне значення нормованого значення його □□ □□□□□□□□. Ми пропонуємо, що висновки до аналізу аспектів мають включати діаграми та таблицю, яка показує фактичні значення □□□□□□□, □□□□□□□ та відні значення для кожного виду аспекту.

3.4 Архітектура системи

Система була розроблена з використанням парадигми модульного програмування. На рисунку 3.3 показана запропонована архітектура. Модуль збору даних (DCM) відповідає за отримання думки з набору даних веб-джерел. Цей модуль складається з певної кількості веб-сканерів, які специфічні для кожного джерела. Сканери аналізують веб-сторінки, що містять відгуки, й попередньо оброблюють результати, створюючи файли, розділені комами, що містять завантажені документи.

Модуль видобутку думки (ОММ) реалізує запропоновані алгоритми на основі аспектів для виявлення думок на заданому наборі документів. Кожен документ, що зберігається, розділений на речення, які потім розділяються на елементи; потім застосовуються пошук частин мови POS та синтаксичні обробки. Необхідно виконати два різні завдання, визначення аспектів та визначення орієнтації, для яких входять два підмодулі:

1) Підмодуль визначення аспектів: відповідає за застосування алгоритму визначення аспекту до набору POS-мічених речень. Як ми вже говорили, цей алгоритм базується на [24], в якому використовуються найчастіші іменники та ІФ для вилучення аспектів.
2) Підмодуль визначення орієнтації: цей підмодуль застосовує алгоритми, наведені в розділі 3.3, для визначення орієнтації висновку щодо даного аспекту. Він також витягує набір прикметників, що з'являються поруч із кожним аспектом.

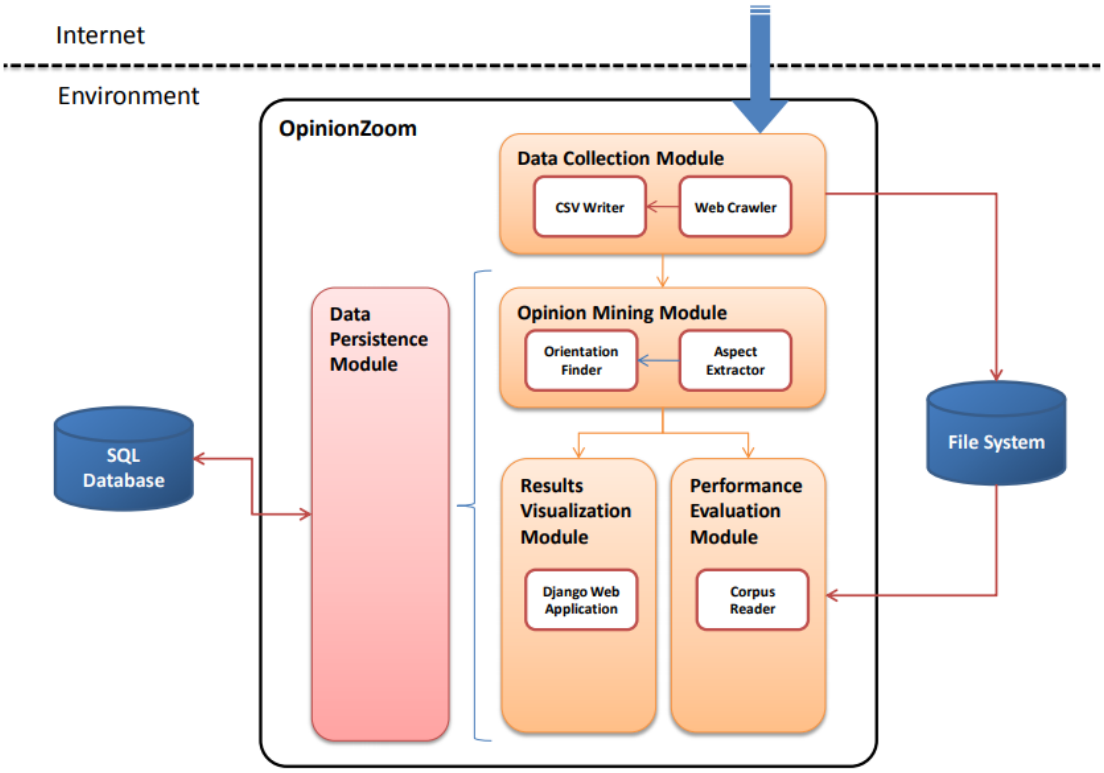

Рисунок 3.3 - Архітектура систуми аналізу неструктурованих текстових даних

Модуль візуалізації результатів (RVM) - це видима частина програми яка безпосередньо взаємодіє з користувачем. Користувачі можуть надавати дані про думки в систему, які потім можуть бути використані для застосування процесу виявлення думок. Результати включають такі функції:

- Підсумки на основі аспектів: стрічкові діаграми, в яких кожен рядок вимірює кількість позитивних та негативних згадок кожного атрибута або компонента одного продукту. Стрічки сортуються відповідно до відносного значення.
- Бульбашкова діаграма прикметників: найближчі прикметники у всіх реченнях, де один аспект з'являється, відображаються на бульбашковій діаграмі. Розмір кожної бульбашки пропорційний

кількості разів, скільки кожний прикметник був використаний для опису аспекту.

- Оригінальні висновки: список всіх оригінальних речень також відображається одночасно, відокремлюючи з них позитивні або негативні.

Система також забезпечує інтерфейс тегів, який допомагає користувачам витягувати думки з документів та змінювати результати алгоритму у випадку неправильної класифікації. Крім того, після застосування алгоритмів пошуку думок, система пропонує інтерфейс, який дозволяє користувачам переглядати список отриманих аспектів та вибрати ті, які він дійсно хоче зберегти. Ці дві функціональні можливості дозволяють отримати відгук про відповідність від кінцевих користувачів. Таким чином, вибір та операції, що виконуються користувачами, зберігаються, а потім використовуються для покращення системи.

Модуль оцінювання продуктивності (PEM) відповідає за отримання набору індексів, що оцінюють ефективність алгоритмів пошуку думок. Для цього система дозволяє користувачам розробляти, а потім надавати спеціально анотовані тіла, дотримуючись структури, яка відображається на рисунку 3.4. Для полегшення процесу анотації пропонуються також рекомендації та приклади. У результаті три завдання можна оцінити шляхом порівняння результатів процесу вилучення з представленими корпораціями:

- витягнення явного аспекту, для вимірювання ефективності алгоритму вилучення явного аспекту;
- класифікація суб'єктивності для оцінки ефективності думки вилучення речення;
- класифікація почуттів, для вимірювання точності прогнозування орієнтації кожного аспектного виразу в кожному реченні (□□, □) для позитивного класу. Сервіс, наданий цим модулем, має

вирішальне значення для розуміння корисності системи в певній темі або домені. Це є важливою відмінністю між реалізованим та іншими існуючими інструментами.

Нарешті, модуль збереження даних або DPM управляє всіма операціями бази даних і становить модельний рівень для всієї системи. Рівень даних реалізується за допомогою двох реляційних моделей, які підтримують всі дані, які необхідно зберігати.

3.5 Експерименти та промисловість застосування

У цьому розділі показана прикладну програму, де запропонована архітектура була реалізована за допомогою Python та JavaScript. Також використовували нашу програму та отримані нами дані для створення наборів даних (лінгвістичних корпусів) для оцінки ефективності алгоритмів виявлення думок, що реалізуються в OMM. Дослідження використовувало бібліотеки NLTK6 для завдань NLP в OMM та express для RVM.

line	
1	*[c1][s1]* place[+], comfort[+][u], location[+][u] ### good place to stay at the end of a long flight in that it is very comfortable, with many facilities, and in the town , by the shore .
2	*[c1][s2]* ### however, puerto montt is not the best of places to explore.
3	*[c1][s3]* ### a better place is puerto varas which is just as near to the airport and has far more attractions.
4	*[c1][s4]* ### could not fault this aparthotel.
5	*[c2][s1]* hotel[+] ### my fiance and i spent three nights here in march 2012 and it's a sweet, quaint hotel.
6	*[c2][s2]* reservation[-] ### with that said, i called in early february 2012 to make a reservation and it got lost/misplaced.

Рисунок 3.4 - Структура анотованого документу

3.5.1 Оцінка продуктивності алгоритму

У першу чергу, використовуючи DCM, відгуки про навчання були завантажені в систему, які були перекладені англійською мовою. Згодом відгуки були збережені у двох різних файлах CSV. Для того, щоб згенерувати анотований корпус для оцінки ефективності алгоритмів, випадковим чином обираються 100 відгуків. Пізніше кожний огляд був розділений на елементи у речення за допомогою алгоритму безконтрольного машинного навчання. Нарешті, кожен огляд був анотований вручну. Речення, які здавалися неоднозначними чи дійсно важкими для позначення відкидалися.

У таблиці 3.1 наведено загальний опис сформованих тіл. В обох випадках близько 78% речень містять думки. Тим не менш, як і очікувалося, є значна кількість речень без вираження думок, які вносять шум в процес визначення.

Таблиця 3.1 - Деталі відгуків

Кількість відгуків	100
Загальна кількість речень	487
Кількість речень з думками	376
Загальна кількість/кількість з думками	77,21%

У таблиці 3.2 наведено відомості про вирази аспектів, які були визначені вручну. Слідуючи згаданим вище позначенням, ті вирази, які виражаються у вигляді іменників або ІФ у реченні з явним аспектним

виразом та неявних аспектів у всіх інших випадках. Результати показують, що явні аспектні вирази є найпоширенішими, що становить близько 77% усіх визначених виразів. Коли деякі аспекти виразів з'являються як явним, так і неявним чином, вони вважаються явними. З іншого боку, вирази визначеного аспекту, які є суто неявними, наявні у великій кількості та в обох випадках становить майже 20%. Простий перегляд показав, що більшість з цих аспектів були позначені прикметниками.

Подальший аналіз набору даних полягав у знаходженні найкращого розподілу кількості речень для кожного випадку. На рисунку 3.5 показані діаграми з найбільш відповідними дискретними розподілами та їх параметрами, отриманими з використанням оцінки максимальної правдоподібності (MLE). Як видно, для обох випадків оптимальним розподілом було негативне біноміальне. Той факт, що параметри EMV були дещо різними для кожного випадку, відповідали відмінностям у кількості речень для кожного випадку (див. Таблицю 3.1), яку ми спостерігали в першому аналізі.

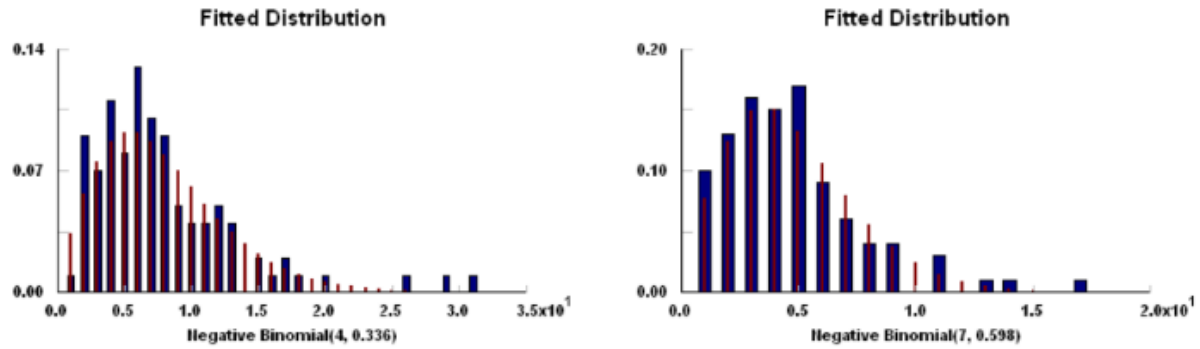

Рисунок 3.5 - Діаграми з найбільш відповідними дискретними розподілами

Таблиця 3.2 Аспекти визначені вручну

Тип аспекту	Відгуки

	Кількість	Відношення
Явний	229	73,87%
Явний і неявний	30	9,68%
Неявний	51	16,45%
Загальна кількість	310	100%

Ми впровадили всі специфікації PEM, а потім оцінили, як запропоновані алгоритми виявлення думок виконуються при застосуванні до відгуків про навчання з використанням отриманого ядра. Тут представлена найкраща загальна продуктивність, отримана шляхом аналізу чутливості щодо найбільш чутливого параметра - мінімального правила підтримки для вилучення виразів аспектів, як це визначено у [24]. Precision, Recall та F-measure були розраховані для шести різних значень цього параметра для кожного завдання. Згодом за допомогою F-міри була обрана найкраща модель. У таблиці 3.3 наведені отримані значення.

Таблиця 3.3 - Оптимальні значення

Індекс	P	R	F-m
Явно отримані аспекти	33%	29%	31%
Суб’єктивна класифікація	79%	93%	85%
Класифікація настрою	89%	93%	91%

Ці результати показують, що продуктивність за завданням визначення аспектів є досить низькою в області навчання. Високий відсоток визначених виразів не відповідає вираженню реальних аспектів. З іншого боку, класифікація настроїв демонструє досить хороші результати, але в цьому випадку більшість можливих висновків складно довести,

оскільки це завдання оцінювалось лише для виражених аспектів. Оскільки ці вирази є дещо простішими, задача визначення сентиментальної орієнтації на них також спрощується. Отже, точність(P) і відкликання(R) можуть зменшуватися, коли розглядаються всі аспектні вирази.

Результати також підтверджують властивості оглядів сервісів. Історії, розказані авторами відгуків, містять в собі об'єкти, що не відповідають атрибутам або компонентам сервісу, є причиною низької точності, отриманою для завдання явного вилучення аспекту. Крім того, іменники та набори ІФ, які не зустрічаються з високими відносними частотами, можливо, потребують певного спеціального режиму, щоб їх було витягнуто, маючи на увазі, що багато виразів можна використовувати для позначення одного і того ж аспекту. У [24] автори запропонували метод визначення цих нечастотних аспектних виразів шляхом експлуатації їхніх зв'язків з частими словами думки. У цій роботі цей метод не враховувався, оскільки у випадку [24] визначення нечіткі аспектні вирази лише означали покращення у 15% для відкликання, за рахунок зменшення точності майже на 7%. Однак, зважаючи на погані результати, які були отримані, цікаво оцінити, як цей крок покращить або погіршить ефективність у даному випадку. З іншого боку, як зазначено в [24], причиною, яка, ймовірно, пояснює точність, яка є трохи нижчою, ніж у завдання класифікації суб'єктивності є той факт, що у рецензуванні навчального продукту є багато речень, що не висвітлюються. Оскільки алгоритм позначає деякі з цих речень як речення з думками, бо вони містять як вирази аспектів товару, так і деякі слова з лексикону думок, точність зменшується.

3.5.2. Порівняння з підходом Лю

З результатів, наведених у таблиці 3.4, можна помітити важливе покращення щодо завдання вилучення суб'єктивних речень. У випадку Лю, середній відклик висновку висловлювання думок становить майже 70%, а середня точність цього самого завдання становить 64%. Тут, хоча точність збільшилася на 10%, найважливішим поліпшенням є відкликання, в цьому випадку на 25% вище. З іншого боку, класифікація почуттів показує покращення, що вище, ніж у випадку Лю. Нарешті, той факт, що завдання з вилучення аспекту призводить до поганих результатів, причому, з використанням невиправданого підходу Лю на підході до відгуків про туристичні продукти, зменшення майже на 40% підтверджує, що особливості, які ми знайшли в області, повинні бути спеціально розглянуті для отримання гарних результатів

Таблиця 3.4 - Середня продуктивність

Індекс	Точність (P)		Відклик (R)		F-міра	
	тут	Б. Лю	тут	Б. Лю	тут	Б. Лю
Явно отримані аспекти	33%	69%	29%	59%	31%	64%
Суб’єктивна класифікація	79%	63%	93%	71%	85%	67%
Класифікація настрою	89%	90%	93%	90%	91%	90%

3.5.3. Оцінка підсумовування

Оскільки ми маємо намір показувати витягнуті аспекти для користувачів, важливо також оцінити, як працює RVM. Веб додаток, розроблений з використанням express з RVM показує підсумки на основі аспектів, в яких таблиця показує фактичні значення Позитивного балу, Негативний показник та відносне значення для кожного аспектного виразу. На рисунку 3.6 наведено приклад. Натискаючи назву кожного стовпця, таблиця та стрічкова діаграма сортуються відповідно до натиснутого стовпця (кожен клік чергується між зростанням або зменшенням сортування.) Натискаючи одне вираження аспекту, користувач переходить на сторінку, де відображаються конкретні відомості про це. На цих сторінках відображаються відповідні бульбашкові графіки прикметників, побудовані з використанням двох найближчих прикметників. Як видно на рисунку 3.7, діаграма дійсно пропонує цінну інформацію, яка вказує на те, що студенти схильні описувати навчання з використанням сильних позитивних прикметників, таких як приємний, чудовий і красивий. Як згадувалося раніше, RVM також пропонує користувачам інтерфейс для вибору аспектів, які необхідно зберегти; На рисунку 6 показано, як виглядає цей інтерфейс. Зважаючи на низьку продуктивність у задачі з вилучення аспектів, ця функціональність стала важливою у цьому конкретному випадку. Для оцінки ми спочатку розглянемо проблему вилучення аспекту з точки зору пошуку інформації та вимірює точність при k визначених аспектів відповідно до їх відносного значення. Оскільки завдання класифікації почуттів має досить високу продуктивність, ми маємо емпіричні докази

того, що діаграми, що показують вирази з найкращим аспектом k, забезпечують точну та правдиву інформацію для користувачів.

Висновки до розділу

У цій роботі була представлена загальна архітектура системи визначення відношення до аспектів, яка вирішує бізнес-задачі. Основою запропонованого підходу є аспектно-орієнтована техніка, що була вперше запропонована Бін Лю.

Ця робота успішно розширила існуючий метод, щоб узагальнити його до сфери послуг. Внаслідок нових та більш складних правил на основі NLP, модифікація здатна працювати краще, ніж модель Лю, що покращує точність та відкликання згаданих завдань. Ефективність цих правил показує, що характеристики, специфічні до домену, включають численні згадки про продукт або наявність великої кількості речень, що не містять жодних поглядів, є точною характеристикою домену. Це підкреслює важливість врахування особливостей, що існують у галузі.

ПЕРЕЛІК ПОСИЛАНЬ

1. Big Data in Big Companies [Електронний ресурс]. – Режим доступу: http://www.datascienceassn.org/sites/default/files/Big%20Data%20in%20Big%20Companies%20-%20Tom%20Davenport.pdf
2. Big Data Executive Survey 2017 [Електронний ресурс]. – Режим доступу: http://newvantage.com/wp-content/uploads/2017/01/Big-Data-Executive-Survey-2017-Executive-Summary.pdf
3. 37 Big Data Case Studies with Big Results [Електронний ресурс]. – Режим доступу: https://www.businessesgrow.com/2016/12/06/big-data-case-studies/
4. The Value of Big Data and the Internet of Things to the UK Economy [Електронний ресурс]. – Режим доступу: https://www.sas.com/content/dam/SAS/en_gb/doc/analystreport/cebr-value-of-big-data.pdf
5. Big Data Will Revolutionize Learning [Електронний ресурс]. – Режим доступу: https://datafloq.com/read/big-data-will-revolutionize-learning/206
6. Recognizing contextual polarity in phrase-level sentiment analysis [Електронний ресурс]. – Режим доступу: https://people.cs.pitt.edu/~wiebe/pubs/papers/emnlp05polarity.pdf
7. Making objective decisions from subjective data: detecting irony in customer reviews [Електронний ресурс]. – Режим доступу: https://www.tib.eu/en/search/id/elsevier%3Adoi~10.1016%252Fj.dss.2012.05.027/Making-objective-decisions-from-subjective-data/

8. Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers [Електронний ресурс]. – Режим доступу: http://www.aclweb.org/anthology/N15-1057

9. Classification of text documents [Електронний ресурс]. – Режим доступу: https://www.cs.uic.edu/~liub/S-EM/unlabelled.pdf

10. Generalizing Syntactic Structures for Product Attribute Candidate Extraction [Електронний ресурс]. – Режим доступу: http://www.aclweb.org/anthology/N10-1059

11. Induction of decision trees [Електронний ресурс]. – Режим доступу: http://hunch.net/~coms-4771/quinlan.pdf

12. Automatic text categorization by unsupervised learning [Електронний ресурс]. – Режим доступу: http://www.aclweb.org/anthology/C00-1066

13. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S0957417412013267

14. That is your evidence?: Classifying stance in online political debate [Електронний ресурс]. – Режим доступу: https://users.soe.ucsc.edu/~maw/papers/wassa_article.pdf

15. DASA: dissatisfaction-oriented advertising based on sentiment analysis [Електронний ресурс]. – Режим доступу: https://www.researchgate.net/publication/223947318_DASA_Dissatisfaction-oriented_Advertising_based_on_Sentiment_Analysis

16. Long autonomy or long delay?' The importance of domain in opinion mining [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S0957417412012729

17. Manipulation of online reviews: an analysis of ratings, readability, and sentiments [Електронний ресурс]. – Режим доступу: https://www.researchgate.net/publication/220195985_Manipulation_of_online_reviews_An_analysis_of_ratings_readability_and_sentiments

18. Exploring determinants of voting for the "helpfulness" of online user reviews: a text mining approach [Електронний ресурс]. – Режим доступу: https://www.researchgate.net/publication/306069346_Predicting_the_helpfulness_of_online_consumer_reviews

19. Producing high-dimensional semantic spaces from lexical co-occurrence [Електронний ресурс]. – Режим доступу: https://link.springer.com/content/pdf/10.3758%2FBF03204766.pdf

20. A lexicon model for deep sentiment analysis and opinion mining applications [Електронний ресурс]. – Режим доступу: http://kt.ijs.si/markodebeljak/Lectures/Seminar_MPS/2012_on/Seminars_2015_16/Simon%20Brmez/Bibliography/%5B18%5D%20A%20lexicon%20model%20for%20deep%20sentiment%20analysis%20and%20opinion%20mining%20applications.pdf

21. Weakness finder: find product weakness from Chinese reviews by using aspects based sentiment analysis [Електронний ресурс]. – Режим доступу: https://www.researchgate.net/publication/257404010_Weakness_Finder_

Find_product_weakness_from_Chinese_reviews_by_using_aspects_based_sentiment_analysis

22. A new approach to consumer theory. The journal of political economy [Електронний ресурс]. – Режим доступу: https://www.journals.uchicago.edu/doi/abs/10.1086/259131

23. Web data mining: exploring hyperlinks, contents, and usage data [Електронний ресурс]. – Режим доступу: http://sirius.cs.put.poznan.pl/~inf89721/Seminarium/Web_Data_Mining__2nd_Edition__Exploring_Hyperlinks__Contents__and_Usage_Data.pdf

24. Mining opinion features in customer reviews [Електронний ресурс]. – Режим доступу: https://pdfs.semanticscholar.org/ee6c/726b55c66d4c222556cfae62a4eb69aa86b7.pdf

25. A holistic lexicon-based approach to opinion mining [Електронний ресурс]. – Режим доступу: https://www.cs.uic.edu/~liub/FBS/opinion-mining-final-WSDM.pdf